SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 1, pp. 1-37, January 1991

001

APPROXIMATION THEORY FOR LINEAR-QUADRATIC-GAUSSIAN OPTIMAL CONTROL OF FLEXIBLE STRUCTURES*

J. S. GIBSON† AND A. ADAMIAN†

Abstract. This paper presents approximation theory for the linear-quadratic-Gaussian optimal control problem for flexible structures whose distributed models have bounded input and output operators. The main purpose of the theory is to guide the design of finite-dimensional compensators that approximate closely the optimal compensator, which is infinite-dimensional. Design of the optimal compensator separates into an optimal linear-quadratic control problem and a dual optimal state estimation problem; the solution to each problem lies in the solution to an infinite-dimensional Riccati operator equation. The approximation scheme in the paper approximates the infinite-dimensional LQG problem with a sequence of finite-dimensional LQG problems defined for a sequence of finite-dimensional, usually finite-element or modal, approximations of the distributed model of the structure. Two Riccati matrix equations determine the solution to each approximating problem.

The finite-dimensional equations for numerical approximation are developed, including formulas for converting matrix control and estimator gains to their functional representation to allow comparison of gains based on different orders of approximation. Convergence of the approximating control and estimator gains and of the corresponding finite-dimensional compensators is studied. Also, convergence and stability of the closed-loop systems produced with the finite-dimensional compensators are discussed. The convergence theory is based on the convergence of the solutions of the finite-dimensional Riccati equations to the solutions of the infinite-dimensional Riccati equations. A numerical example with a flexible beam, a rotating rigid body, and a lumped mass is given.

Key words. linear-quadratic-Gaussian optimal control, approximation theory, flexible structures, distributed systems

1. Introduction.

The first question that must be answered when designing a controller for a flexible structure is whether a finite-dimensional model is sufficient as a basis for a controller that will produce the required performance, or is a distributed model necessary? While some structures can be modeled well by a fixed number of dominant modes, there are structures whose flexible character can be captured sufficiently for precise control only by a distributed model. Still others—perhaps most of the aerospace structures of the future—can be modeled sufficiently for control purposes by some finite-dimensional approximation, but an adequate approximation may be impossible to determine before design of the controller, or compensator. This paper deals with structures that are flexible enough to require a distributed model in the design of an optimal LQG compensator.

The linear-quadratic-Gaussian optimal control problem for distributed, or infinite-dimensional, systems is a generalization to Hilbert space of the LQG problem for finite-dimensional systems. The solution to the infinite-dimensional problem yields an infinite-dimensional state-estimator-based compensator, which is optimal in the context of this paper. By a separation principle [B1], [C4], the problem reduces to a deterministic linear-quadratic optimal control problem an an optimal estimation, or filtering, problem with Gaussian white noise. The solutions to both the control and filtering problems involve Riccati operator equations, which are generalizations of the Riccati matrix equations in the finite-dimensional case. Current results on the infinite-dimensional LQG problem are most complete for problems where the input and measurement operators are bounded, as this paper requires throughout. This bounded-

* Received by the editors February 24, 1986; accepted for publication (in revised form) May 26, 1989.

† Department of Mechanical, Aerospace, and Nuclear Engineering, University of California, Los Angeles, California 90024. This research was supported by Air Force Office of Scientific Research grant 840309 and by grant 957114 from the Jet Propulsion Laboratory, Pasadena, CA.

ness also permits the strongest approximation results here. For related control problems with unbounded input and measurement, see [C3], [C5], [C6], [D2], [I2], [I3], [L1]-[L4].

Our primary objective in this paper is to approximate the optimal infinite-dimensional LQG compensator for a distributed model of a flexible structure with finite-dimensional compensators based on approximations to the structure, and to have these finite-dimensional compensators produce near optimal performance of the closed-loop system. We discuss how the gains that determine the finite-dimensional compensators converge to the gains that determine the infinite-dimensional compensator, and we examine the sense in which the finite-dimensional compensators converge to the infinite-dimensional compensator. With this analysis, we can predict the performance of the closed-loop system consisting of the distributed plant and a finite-dimensional compensator that approximates the infinite-dimensional compensator.

Our design philosophy is to let the convergence of the finite-dimensional compensators indicate the order of the compensator that is required to produce the desired performance of the structure. The two main factors that govern rate of convergence are the desired performance (e.g., fast response) and the structural damping. We should note that any one of our compensators whose order is not sufficient to approximate the infinite-dimensional compensator closely will not, in general, be the optimal compensator of that fixed order, i.e., the optimal fixed-order compensator that would be constructed with the design philosophy in [B7], [B8]. But as we increase the order of approximation to obtain convergence, our finite-dimensional compensators become essentially identical to the compensator that is optimal over compensators of all orders.

An important question, of course, is how large a finite-dimensional compensator we must use to approximate the infinite-dimensional compensator. In [G6]-[G8], and [M1], we have found that our complete design strategy yields compensators of reasonable size for distributed models of complex space structures. This strategy in general requires two steps to obtain an implementable compensator that is essentially identical to the optimal infinite-dimensional compensator: the first step determines the optimal compensator by letting the finite-dimensional compensators converge to it, the second step reduces the order of a large (converged) approximation to the optimal compensator. The first step, which involves control theory and approximation theory for distributed systems, is the subject of this paper. For the second step, a simple modal truncation of the large compensator sometimes is sufficient, but there are more sophisticated methods in finite-dimensional control theory for order reduction. For example [G9], [M1], we have found that balanced realizations [M2] work well for reducing large compensators.

The approximation theory in this paper follows from the application of approximation results in [B6], [G3], [G4] to a sequence of finite-dimensional optimal LQG problems based on a Ritz–Galerkin approximation of the flexible structure. For the optimal linear-quadratic control problem, the approximation theory here is a substantial improvement over that in [G1] because here we allow rigid-body modes, more general structural damping (including damping in the boundary), and much more general finite-element approximations. These generalizations are necessary to accommodate common features of complex space structures and the most useful finite-element schemes. For example, we write the equations for constructing the approximating control and estimator gains and finite-dimensional compensators in terms of matrices that are built directly from typical mass, stiffness and damping matrices for flexible structures, along with actuator influence matrices and measurement matrices.

For the estimator problem, this paper presents some of the first rigorous approximation theory. (We have used less complete versions of the results in previous research [G7]-[G9], [M1].) We note also [I1] that as in the finite-dimensional case, the infinite-dimensional optimal estimation problem is the dual of the infinite-dimensional optimal control problem, and the solutions to both problems have the same structure. Because we exploit this duality to obtain the approximation theory for the estimation problem from the approximation theory for the optimal control problem, the analysis in this paper is almost entirely deterministic. We discuss the stochastic interpretation of the estimation problem and the approximating state estimators briefly, but we are concerned mainly with deterministic questions about the structure and convergence of approximations to an infinite-dimensional compensator and the performance—especially stability—of the closed-loop systems produced by the approximating compensators.

The paper has two main parts, which correspond roughly to the separation of the optimal LQG problem into an optimal linear-quadratic regulator problem and an optimal state estimation problem. The first half, §§ 2-6, deal with the control system and the optimal regulator problem. Sections 7-10 treat the state estimator and the compensator that is formed by applying the control law of the first half of the paper to the output of the estimator.

While this paper is primarily theoretical, we present a detailed numerical example in §§ 6 and 10. The structure in this example consists of an Euler-Bernoulli beam attached to a rotating rigid hub on one end and to a lumped mass on the other end. We emphasize the fact that we do not solve, or even write down, the coupled partial and ordinary differential equations of motion. For both the definition and numerical solution of the problem, only the kinetic and strain energy functionals and a dissipation functional for the damping are required. In § 6, we show the approximating functional control gains obtained by using a standard finite-element approximation of the beam, and we discuss the effect on convergence of structural damping and of the ratio of state weighting to control weighting in the performance index. As suggested by a theorem in § 5, the functional gains do not converge when no structural damping is modeled.

In § 10, we complete the compensator design for the example. Assuming that white noise corrupts the single measurement and that distributed white noise disturbs the structure, we compute the gains for the finite-dimensional estimators and show the functional estimator gains. As in the control problem, the functional gains do not converge when no damping is modeled. We apply the control laws computed in § 6 to the output of the estimators in § 10 to construct the finite-dimensional compensators, and we show the frequency response of these compensators. As predicted by § 9.3, the frequency response of the nth compensator converges to the frequency response of the optimal infinite-dimensional compensator as n increases. In § 10.3, we discuss the structure and dimension of the finite-dimensional compensator that should be implemented.

2. The control system.

We consider the system

$$\ddot{x}(t)+D_0\dot{x}(t)+A_0x(t)=B_0u(t), \qquad t>0 \tag{2.1}$$

where $x(t)$ is in a real Hilbert space H and $u(t)$ is in R^m for some finite m. The linear stiffness operator A_0 is densely defined and self-adjoint with compact resolvent and has at most a finite number of negative eigenvalues. We will postpone discussion of the damping operator D_0 momentarily, except to say that it is symmetric and nonnegative. The input operator B_0 is a linear operator from R^m to H, and hence is bounded.

By natural modes, we will mean the eigenvectors ϕ_j of the eigenvalue problem

$$\lambda_j\phi_j = A_0\phi_j. \tag{2.2}$$

From our hypotheses on A_0, we know that these eigenvalues form an infinitely increasing sequence of real numbers of which all but a finite number are positive. Also, the corresponding eigenvectors are complete in H and satisfy

$$\langle\phi_i, \phi_j\rangle_H = \langle A_0\phi_i, \phi_j\rangle_H = 0, \qquad i \neq j. \tag{2.3}$$

For $\lambda_j > 0$, $\omega_j = \sqrt{\lambda_j}$ is a *natural frequency*.

Remark 2.1. Our analysis includes the system

$$M_0\ddot{x}(t) + D_0\dot{x}(t) + A_0x(t) = B_0u(t), \qquad t > 0 \tag{2.1'}$$

where the mass operator M_0 is a self-adjoint, bounded, and coercive linear operator on a real Hilbert space H_0. The operators A_0, B_0, and D_0 in (2.1') have the same properties with respect to H_0 that the corresponding operators in (2.1) have with respect to H. To include (2.1') in our analysis, we need only take H to be H_0 with the norm-equivalent inner product $\langle\cdot,\cdot\rangle_H = \langle M_0\cdot,\cdot\rangle_{H_0}$, and multiply (2.1') on the left by M_0^{-1}. In H, the operator $M_0^{-1}A_0$ is self-adjoint with compact resolvent, and $M_0^{-1}D_0$ is symmetric and nonnegative. With no loss of generality then, we will refer henceforth only to (2.1) and assume that the H-inner product accounts for the mass distribution.

2.1. The energy spaces and the first-order form of the system.

2.1.1. The elastic-strain-energy space V and total-energy space E. We choose a bounded, self-adjoint linear operator A_1 on H such that $\tilde{A}_0 = A_0 + A_1$ is coercive; i.e., there exists $\rho > 0$ for which

$$\langle\tilde{A}_0x, x\rangle_H \geqq \rho\|x\|_H^2, \qquad x \in D(\tilde{A}_0) = D(A_0). \tag{2.4}$$

In applications such as our example in § 6, it is natural to select for A_1 an operator whose null space is the orthogonal complement (in H) of the eigenspace of A_0 corresponding to nonpositive eigenvalues. Obviously, any A_1 that makes $\tilde{A}_0$ coercive must be positive definite on the nonpositive eigenspace of A_0.

With A_1 chosen, we define the Hilbert space V to be the completion of $D(A_0)$ with respect to the inner product $\langle v_1, v_2\rangle_V = \langle\tilde{A}_0v_1, v_2\rangle_H$, v_1 and $v_2 \in D(A_0)$. Note that $V = D(\tilde{A}_0^{1/2})$ and $\langle v_1, v_2\rangle_V = \langle\tilde{A}_0^{1/2}v_1, \tilde{A}_0^{1/2}v_2\rangle_H$. (Since A_1 is a bounded operator on H, different choices of A_1 yield V's with equivalent norms, thus containing the same elements.)

In the usual way, we will use the imbedding

$$V \subset H = H' \subset V'$$

where the injections from V into H and from H into V' are continuous with dense ranges. We denote by Λ_V the Riesz map from V onto its dual V'. Then $\tilde{A}_0$ is the restriction of Λ_V to $D(A_0)$ in the sense that

$$(\Lambda_Vv_1)v = \langle v, \tilde{A}_0v_1\rangle_H, \qquad v_1 \in D(A_0), \quad v \in V. \tag{2.5}$$

Now we define the *total energy space* $E = V \times H$, noting that when A_0 is coercive and $x(t)$ is the solution to (2.1), then $\|(x(t), \dot{x}(t))\|_E^2$ is twice the total energy (kinetic plus potential) in the system. We want to write (2.1) as a first-order evolution equation on E. To do this, we must determine the semigroup generator for the open-loop system. We will derive this generator by constructing its inverse explicitly. This approach seems mathematically efficient, and we will need the inverse of the generator for the approximation scheme. First, we must state our precise hypotheses on damping and discuss its representation.

2.1.2. The damping functional and operator. Actually, we do *not* require an operator D_0 defined from some subset of H into H. Rather, we assume only that there exists a damping functional

$$d_0(v_1, v_2): V \times V \to R \tag{2.6}$$

such that d_0 is bilinear, symmetric, continuous (on $V \times V$), and nonnegative. Under these hypotheses, there is a unique bounded linear operator Λ_D from V into V' such that

$$d_0(v, v_1) = (\Lambda_D v_1)v, \qquad v_1, v \in V. \tag{2.7}$$

The operator $D_V = (\Lambda_V^{-1}\Lambda_D)$ is then a bounded linear operator from V to V, and D_V is self-adjoint (on V) because d_0 is symmetric. Also

$$d_0(v, v_1) = \langle v, D_V v_1 \rangle_V = \langle D_V v, v_1 \rangle_V, \qquad v_1, v \in V. \tag{2.8}$$

Remark 2.2. We chose to begin our description of the control system model with (2.1) because its form is familiar in the context of flexible structures. In applications such as the example in § 6, however, it is easier to begin with a strain-energy functional from which the correct strain-energy inner product for V is obvious. The stiffness operator is then defined in terms of the Riesz map for V (see [S3] for this approach). Either way, the only thing that needs to be computed in applications is the V-inner product; an explicit A_0 need not be written down.

2.1.3. The semigroup generator. We define $\tilde{A}^{-1} \in L(E, E)$ by

$$\tilde{A}^{-1} = \begin{bmatrix} -D_V & -\tilde{A}_0^{-1} \\ I & 0 \end{bmatrix}. \tag{2.9}$$

This operator is clearly one-to-one, and its range is dense, since V is dense in H and $D(A_0)$ is dense in V. Now, we take

$$\tilde{A} = (\tilde{A}^{-1})^{-1}. \tag{2.10}$$

Direct calculation of the inner product shows

$$\left\langle \tilde{A}^{-1}\begin{bmatrix} v \\ h \end{bmatrix}, \begin{bmatrix} v \\ h \end{bmatrix} \right\rangle_E = -d_0(h, h), \tag{2.11}$$

so that $\tilde{A}$ is dissipative with dense domain. Also, since $D(\tilde{A}^{-1}) = E$, $\tilde{A}$ is maximal dissipative by [G1, Thm. 2.1]. Therefore, $\tilde{A}$ generates a C_0-contraction semigroup on E.

Finally, the open-loop semigroup generator is

$$A = \tilde{A} + \begin{bmatrix} 0 & 0 \\ A_1 & 0 \end{bmatrix}, \qquad D(A) = D(\tilde{A}) \tag{2.12}$$

where A_1 is the bounded linear operator chosen to make $\tilde{A}_0$ coercive. With

$$B = \begin{bmatrix} 0 \\ B_0 \end{bmatrix} \in L(R^m, E), \tag{2.13}$$

the first-order form of (2.1) is

$$\dot{z}(t) = Az(t) + Bu(t), \qquad t > 0 \tag{2.14}$$

where $z = (x, \dot{x}) \in E$.

We should note that Showalter [S3, Chap. VI] elegantly derives a semigroup generator for a class of second-order systems that includes the flexible-structure model here. The presentation here is most useful for our approximation theory because of the explicit construction of the inverse of the semigroup generator.

In many structural applications, the open-loop semigroup is analytic. Showalter obtains an analytic semigroup when the damping functional is V-coercive, for example, when there exists a damping operator D_0 that is both A_0-bounded and as strong as A_0. Such a damping operator results from the Kelvin–Voigt viscoelastic material model. Chen and Russell [C1] have shown that the semigroup is analytic for a class of damping operators involving $A_0^{1/2}$, and recently Chang and Lasiecka [C6] have shown that the semigroup is analytic if $c_0 A_0^{\mu} \leqq D_0 \leqq c_1 A_0^{\mu}$ for some μ such that $\frac{1}{2} \leqq \mu \leqq 1$ and c_0 and c_1 positive and finite. If the damping operator is bounded relative to A_0^{μ} for $\mu < 1$, then A has compact resolvent.

Finally, we can guarantee that the open-loop semigroup generator is a spectral operator (i.e., its eigenvectors are complete in E) only for a damping operator that is a linear combination of an H-bounded operator and a fractional power of A_0. However, nowhere do we use or assume anything about the eigenvectors of either the open-loop or the closed-loop semigroup generator. The natural modes—of undamped free vibration—in (2.2) are always complete in both H and V.

2.2. The adjoint of $\tilde{A}$. Since D_V is self-adjoint on V, direct calculation shows that $\tilde{A}^{-*}$ (i.e., the adjoint of $\tilde{A}^{-1}$ with respect to the E-inner product) is

$$\tilde{A}^{-*} = \begin{bmatrix} -D_V & \tilde{A}_0^{-1} \\ -I & 0 \end{bmatrix}. \tag{2.15}$$

Then $\tilde{A}^* = (\tilde{A}^{-*})^{-1}$. Having $\tilde{A}^{-*}$ explicitly facilitates proving strong convergence for approximating adjoint semigroups.

2.3. Exponential stability. The following theorem states that, if there are no rigid-body modes and if the damping is coercive (basically, all structural components have positive damping), then the open-loop system is uniformly exponentially stable. That the decay rate given depends only on the lower bound for the stiffness operator and the upper and lower bounds for the damping functional is essential for convergence results for the approximating optimal control problems of subsequent sections. The theorem is a generalization of Theorem 6.1 of [G1] to allow more general damping, but the proof is entirely different and much nicer. The current proof uses an explicit Lyapunov functional for the homogeneous part of the system in (2.14). Recall that $T(\cdot)$ is the open-loop semigroup, with generator A, and E is the total energy space $V \times H$.

THEOREM 2.3. *Suppose that A_0 and d_0 are H-coercive. Let ρ be the positive constant in* (2.4), *and let δ_0 and δ_1 be positive constants such that*

$$\delta_0 \|v\|_H^2 \leqq d_0(v, v) \leqq \delta_1 \|v\|_V^2, \qquad v \in V. \tag{2.16}$$

Then

$$\|T(t)\| \leqq \left(1 + \frac{\delta_0}{\sqrt{\rho}} + \frac{\delta_1 \delta_0}{2}\right)^{1/2} \exp\left[-t \Big/ \left(\frac{2}{\delta_0} + \frac{2}{\sqrt{\rho}} + \delta_1\right)\right], \qquad t \geqq 0. \tag{2.17}$$

Proof. For $\gamma > \max\{1/\sqrt{\rho}, 2/\delta_0\}$, define $Q \in L(E)$ as

$$Q = \begin{bmatrix} (\gamma I + \Lambda_V^{-1} \Lambda_D) & A_0^{-1} \\ I & \gamma I \end{bmatrix}. \tag{2.18}$$

Since D_V is self-adjoint and nonnegative on V, Q is self-adjoint and coercive on E. It is shown in [G6] that $\langle QAz, z \rangle_E \leqq -\|z\|_E^2$ and that (2.17) holds. □

3. The optimal control problem. Section 3.1 presents some preliminary definitions and results for the optimal linear-quadratic regulator problem on an arbitrary real Hilbert space. These results are generic in the sense that the Hilbert space E is not necessarily the energy space of § 2, and the operators A, B, etc., do not necessarily represent an abstract flexible structure as in § 2. In the second half of the paper, having such generic results will allow us to obtain the approximation theory for the infinite-dimensional state estimator from the analogous results for the control problem. Section 3.2 gives some important implications of the general results for the case where the control system is that defined in § 2. The proofs of the new results in this section (those after Theorem 3.3) are technical and sometimes tedious. They are given in [G6].

3.1. The generic optimal regulator problem. Let a linear operator A generate a C_0-semigroup $T(t)$ on a real Hilbert space E, and suppose $B\in L(R^m, E)$, $Q\in L(E, E)$, and $R\in L(R^m)$, with Q nonnegative and self-adjoint and R positive definite and symmetric. The *optimal control problem on* E is to choose the control $u\in L_2(0,\infty; R^m)$ to minimize the cost functional

$$J(z(0), u)=\int_0^\infty (\langle Qz(t), z(t)\rangle_E+\langle Ru(t), u(t)\rangle_{R^m})\,dt \tag{3.1}$$

where the state $z(t)$ is given by

$$z(t)=T(t)z(0)+\int_0^t T(t-\eta)Bu(\eta)\,d\eta, \qquad t\geqq 0. \tag{3.2}$$

DEFINITION 3.1. A function $u\in L_2(0,\infty; U)$ is an *admissible control for the initial state* z, or simply an *admissible control for* z, if $J(z, u)$ is finite; i.e., if the state $z(t)$ corresponding to the control $u(t)$ and the initial condition $z(0)=z$ is in $L_2(0,\infty; E)$.

DEFINITION 3.2. Let the operators A, B, Q, and R be as defined above. An operator Π in $L(E)$ is a solution of the Riccati algebraic equation if Π maps the domain of A into the domain of A^* and satisfies the Riccati algebraic equation

$$A^*\Pi+\Pi A-\Pi BR^{-1}B^*\Pi+Q=0. \tag{3.3}$$

THEOREM 3.3 (Theorems 4.6 and 4.11 of [G4]). *There exists a nonnegative self-adjoint solution of the Riccati algebraic equation if and only if, for each* $z\in E$, *there is an admissible control for the initial state* z. *If* Π *is the minimal nonnegative self-adjoint solution of* (3.3), *then the unique control* $u(\cdot)$ *that minimizes* $J(z, u)$ *and the corresponding optimal trajectory* $z(\cdot)$ *are given by*

$$u(t)=-R^{-1}B^*\Pi z(t) \tag{3.4}$$

and

$$z(t)=S(t)z \tag{3.5}$$

where $S(t)$ *is the semigroup generated by* $A-BR^{-1}B^*\Pi$. *Also,*

$$J(z, u)=\min_v J(z, v)=\langle \Pi z, z\rangle_E. \tag{3.6}$$

If, for each initial state and admissible control,

$$\lim_{t\to\infty}\|z(t)\|=0, \tag{3.7}$$

then there exists at most one nonnegative self-adjoint solution of (3.3). *If* Q *is coercive, then* (3.7) *holds for each initial state and admissible control and* $S(t)$ *is uniformly exponentially stable.*

We will refer to $T(t)$ as the *open-loop semigroup* and to $S(t)$ as the *optimal closed-loop semigroup.*

To prepare for the convergence analysis in §§ 5 and 9, we now present some estimates for the decay rate of the closed-loop system in the optimal control problem.

THEOREM 3.4. *Suppose that the open-loop semigroup* $T(\cdot)$ *satisfies*

$$\|T(t)\| \leqq M_1 e^{\alpha_1 t}, \qquad t \geqq 0, \tag{3.8}$$

for positive constants M_1 *and* α_1, *that* Π *is the minimal nonnegative self-adjoint solution to* (3.3), *and that* $S(t)$ *is the optimal closed-loop semigroup in Theorem* 3.3. *If there exists a constant* M_0 *such that, for each* $z \in E$,

$$\int_0^\infty \|S(t)z\|^2 \, dt \leqq M_0(\langle \Pi z, z\rangle_E + \|z\|^2) \tag{3.9}$$

and a constant M_0' *such that*

$$\|\Pi\| \leqq M_0', \tag{3.10}$$

then there exist positive constants M_2 *and* α_2, *which are functions of* M_0, M_0', M_1, *and* α_1 *only, such that*

$$\|S(t)\| \leqq M_2 e^{-\alpha_2 t}, \qquad t \geqq 0. \tag{3.11}$$

LEMMA 3.5. *Suppose there exist positive constants* M *and* α *such that*

$$\|T(t)\| \leqq M e^{-\alpha t}, \qquad t \geqq 0. \tag{3.12}$$

If $z(0) \in E$, $h \in L_2(0, \infty; E)$, *and* $z(t) = T(t)z(0) + \int_0^t T(t-s)h(s)\, ds$, *then*

$$\int_0^\infty \|z(t)\|^2 \, dt \leqq \left[\frac{M}{\sqrt{2a}} \|z(0)\| + \frac{M}{\alpha} \|h\|_{L_2} \right]^2. \tag{3.13}$$

LEMMA 3.6. *Suppose that* E *is finite-dimensional and that the pair* (Q, A) *is observable* (*in the usual finite-dimensional sense*). *Then there exists a constant* M, *which is a function of* A, B, *and* Q *only, such that*

$$\int_0^\infty \|z(t)\|^2 \, dt \leqq M \left(\int_0^\infty (\langle Qz(t), z(t)\rangle_E + \|u(t)\|^2) \, dt \right) \tag{3.14}$$

where $z(t)$ *is given by* (3.2).

The next theorem says, among other things, that if the open-loop control system decouples into a finite-dimensional part that is stabilizable (in the usual finite-dimensional sense) and an infinite-dimensional part that is uniformly exponentially stable, then the entire system is uniformly exponentially stabilizable, so that (3.3) has a nonnegative self-adjoint solution.

THEOREM 3.7. *Suppose that there exists a finite-dimensional subspace* $E_0 \subset D(A)$ *such that* E_0 *and* $E_0^\perp$ *reduce* A (*and* $T(t)$), *and write*

$$A = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} \\ B_{21} \end{bmatrix}, \quad Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{12}^* & Q_{22} \end{bmatrix} \tag{3.15}$$

where A_{11} *and* A_{22} *are the restrictions of* A *to* E_0 *and* $D(A) \subset E_0^\perp$, *respectively. Similarly,*

$$T(t) = \begin{bmatrix} T_{11}(t) & 0 \\ 0 & T_{22}(t) \end{bmatrix}. \tag{3.16}$$

Also, suppose that the pair (A_{11}, B_{11}) *is stabilizable and that there exist positive constants* M_1', α_1', *and* β *such that*

$$\|T_{22}(t)\| \leqq M_1' e^{-\alpha_1' t}, \qquad t \geqq 0 \tag{3.17}$$

and

$$\max\{\|B\|, \|Q\|\} \leqq \beta. \tag{3.18}$$

(i) *Then there exists* $F \in L(E, R^m)$ *such that* $A - BF$ *generates a uniformly exponentially stable semigroup on E. Also,* (3.3) *has a nonnegative self-adjoint solution, and the minimal such solution satisfies* (3.10) *with* M_0' *a function of* $A_{11}, B_{11}, R, M_1', \alpha_1'$, *and* β *only.*

(ii) *If* $Q_{12} = 0$ *and the pair* (Q_{11}, A_{11}) *is observable, then there exists a unique nonnegative self-adjoint solution* Π *to* (3.3), *and there exist positive constants* M_2 *and* α_2*—which depend on* $A_{11}, B_{11}, Q_{11}, R, M_1', \alpha_1'$, *and* β *only—such that the optimal closed-loop semigroup satisfies*

$$\|S(t)\| \leqq M_2 e^{-\alpha_2 t}, \qquad t \geqq 0. \tag{3.19}$$

Remark 3.8. When we say in Theorem 3.4 that M_2 and α_2 are functions of M_0, M_0', M_1, α_1 only, we mean, for example, that for two optimal control problems on different spaces E, with different operators A, B, etc., if the same constants M_0, M_0', M_1, and α_1 work in (3.8)-(3.10) for both problems, then the same constants M_2 and α_2 will work in (3.11) for both problems. Similarly, in Theorem 3.7(ii), as long as $E_0, A_{11}, B_{11}, Q_{11}, R, M_1', \alpha_1'$, and β remain the same, the same M_2 and α_2 will work in (3.19) even if $E_0^\perp, A_{22}, B_{21}$, and Q_{22} change.

3.2. Application to optimal control of flexible structures. For the rest of this section, $A_0, A_1, A, T(t), B_0$, and B are the operators defined in § 2.1, and $E = V \times H$ is the energy space defined there.

Remark 3.9. Theorem 3.7 is useful mainly when all but a finite number of modes have coercive damping in the open-loop system and the damped and undamped parts of the open-loop system remain orthogonal. This is the case, for example, with modal damping. The next theorem does not require orthogonality of the damped and undamped parts of the system, but it does require an independent actuator for each undamped mode. The situation of Theorem 3.10 is typical in aerospace structures: any elastic component should have some structural damping, but rigid-body modes are common; for a structure to be controllable, an actuator is required for each rigid-body mode.

THEOREM 3.10. (i) *Suppose that* $A_1 = B_0 B_0^*$ *and that* $\tilde{A}_0 = A_0 + A_1$ *and* $\tilde{d}_0 = d_0 + A_1$ *are H-coercive, so that there exist positive constants* ρ, γ, *and* β *such that, for all* $v \in V$,

$$\|v\|_V^2 \geqq \rho \|v\|_H^2, \tag{3.20}$$

$$\rho \|v\|_H^2 \leqq \tilde{d}_0(v, v) \leqq \gamma \|v\|_V^2, \tag{3.21}$$

and

$$\max\{\|B_0\|, \|Q\|, \|R\|\} \leqq \beta. \tag{3.22}$$

(*The V-continuity of* d_0 *implies the second inequality in* (3.21).) *Then* (3.3) *has a minimal nonnegative self-adjoint solution* Π, *which satisfies* (3.10) *with* M_0' *a function of* ρ, γ, *and* β *only.*

(ii) *Suppose also that*

$$\langle Qz, z\rangle_E \geqq \rho \|z\|_E^2, \qquad z \in E. \tag{3.23}$$

Then the optimal closed-loop semigroup satisfies

$$\|S(t)\| \leqq M_2 e^{-\alpha_2 t}, \qquad t \geqq 0 \tag{3.24}$$

where M_2 and α_2 are positive constants depending on ρ, γ, and β only.

Now we will consider the structure of the optimal control law in more detail. Since $\Pi \in L(E, E)$ and $E = V \times H$, we can write

$$\Pi = \begin{bmatrix} \Pi_0 & \Pi_1 \\ \Pi_1^* & \Pi_2 \end{bmatrix} \tag{3.25}$$

where $\Pi_0 \in L(V, V)$, $\Pi_1 \in L(H, V)$, $\Pi_2 \in L(H, H)$, and Π_0 and Π_2 are nonnegative and self-adjoint. With $z = (x, \dot{x})$, as in § 2, (3.4) becomes

$$u(t) = -R^{-1} B_0^* [\Pi_1^* x(t) + \Pi_2 \dot{x}(t)]. \tag{3.26}$$

Since $B_0 \in L(R^m, H)$, we must have vectors $b_i \in H$, $1 \leqq i \leqq m$, such that

$$B_0 u = \sum_{i=i}^{m} b_i u_i \quad \text{for } u = [u_1 u_2 \cdots u_m]^T \in R^m. \tag{3.27}$$

Also, for $h \in H$,

$$B_0^* h = [\langle b_1, h\rangle_H \langle b_2, h\rangle_H \cdots \langle b_m, h\rangle_H]^T. \tag{3.28}$$

Since $\Pi_1^* x(t)$ and $\Pi_2 \dot{x}(t)$ are elements of H, we see from (3.26) and (3.28) that the components of the optimal control have the feedback form

$$u_i(t) = -\langle f_i, x(t)\rangle_V - \langle g_i, x(t)\rangle_H, \qquad i = 1, \cdots, m \tag{3.29}$$

where $f_i \in V$ and $g_i \in H$ are given by

$$f_i = \sum_{j=1}^{m} (R^{-1})_{ij} \Pi_1 b_j, \quad g_i = \sum_{j=1}^{m} (R^{-1})_{ij} \Pi_2 b_j, \quad i = 1, \cdots, m. \tag{3.30}$$

We call f_i and g_i *functional gains.*

4. The approximation scheme.

4.1. Approximation of the open-loop system.

Hypothesis 4.1. There exists a sequence of finite-dimensional subspaces V_n of V such that the sequence of orthogonal projections P_{Vn} converges V-strongly to the identity where P_{Vn} is the V-projection onto V_n. Also, each V_n is the span of n linearly independent vectors e_j.

Since it should cause no confusion, we will omit the subscript n and write just e_j, keeping in mind that the basis vectors may change from one V_n to another, as in most finite-element schemes. Also, we will refer to the Hilbert space $E_n = V_n \times V_n$, which has the same inner product as $E = V \times H$.

For $n \geqq 1$, we approximate $x(t)$ by

$$x_n(t) = \sum_{j=1}^{n} \xi_j(t) e_j \tag{4.1}$$

where $\xi(t) = (\xi_1(t), \xi_2(t), \cdots, \xi_n(t))^T$ satisfies

$$M^n \ddot{\xi}(t) + D^n \dot{\xi}(t) + K^n \xi(t) = B_0^n u(t), \tag{4.2}$$

and the mass matrix M^n, damping matrix D^n, stiffness matrix K^n, and actuator influence matrix B_0^n are given by

$$\begin{aligned} &M^n=[\langle e_i, e_j\rangle_H], \qquad D^n=[d_0(e_i, e_j)], \\ &K^n=[\langle A_0^{1/2}e_i, A_0^{1/2}e_j\rangle_H]=[\langle e_i, e_j\rangle_V]-[\langle A_1 e_i, e_j\rangle_H], \\ &B_0^n=[\langle e_i, b_j\rangle_H]. \end{aligned} \tag{4.3}$$

Of course, (4.2) can be written as

$$\dot{\eta}=A^n\eta+B^n u \tag{4.2'}$$

where

$$\eta=[\xi^T, \dot{\xi}^T]^T \tag{4.4}$$

and

$$A^n=\begin{bmatrix} 0 & I \\ -M^{-n}K^n & -M^{-n}D^n \end{bmatrix}, \qquad B^n=\begin{bmatrix} 0 \\ M^{-n}B_0^n \end{bmatrix}. \tag{4.5}$$

(*Note.* Throughout this paper, we use the superscript n in the designation of matrices in the nth approximating system and control problem such as A^n, B^n, M^n, etc. Hence the superscript n indicates the order of approximation—it *never* indicates a power of the matrix. By M^{-n}, we denote the inverse of the mass matrix M^n.

In the designation of a linear operator in the nth approximation, we use the subscript n. For example, A_n and B_n are the operators whose matrix representations are A^n and B^n, respectively.)

For convergence analysis, it is useful to note that (4.1) and (4.2) or (4.2′) are equivalent to

$$\dot{z}_n(t)=A_n z_n(t)+B_n u(t) \tag{4.6}$$

where $z_n=(x_n, \dot{x}_n)\in E_n$, and $A_n\in L(E_n)$ and $B_n\in L(R^m, E_n)$ are the operators whose matrix representations are given in (4.5). Also, for any real λ,

$$\lambda-A_n\begin{bmatrix} v_n^1 \\ v_n^2 \end{bmatrix}=\begin{bmatrix} h_n^1 \\ h_n^2 \end{bmatrix} \tag{4.7}$$

is equivalent to

$$(\lambda^2M^n+\lambda D^n+K^n)\alpha^1=(\lambda M^n+D^n)\beta^1+M^n\beta^2 \tag{4.8}$$

and

$$\alpha^2=\lambda\alpha^1-\beta^1 \tag{4.9}$$

if

$$v_n^j=\sum_{i=1}^n \alpha_i^j e_i \quad \text{and} \quad h_n^j=\sum_{i=1}^n \beta_i^j e_i, \quad j=1,2. \tag{4.10}$$

Next, we will prepare to invoke the Trotter-Kato semigroup approximation theorem to show how (4.2), (4.2′), and (4.6) approximate (2.1) and (2.14). First, we will treat the case in which A_0 is coercive (no rigid-body modes), so that $A_1=0$ and $\tilde{A}_0=A_0$; the general case is a straightforward extension. For A_0 coercive, the open-loop semigroup generator A is maximal dissipative. Also, for each n, A_n is dissipative on E_n. The main idea here is to project $(\lambda-A)^{-1}$ onto V_n in a certain inner product and observe that the result is exactly $(\lambda - A_n)^{-1}$, where A_n is the operator on E_n in (4.6) and (4.7). Of course, we need only do this for real $\lambda > 0$.

For real $\lambda > 0$ then, define an inner product on V by

$$\langle \cdot, \cdot \rangle_\lambda = \lambda^2 \langle \cdot, \cdot \rangle_H + \lambda d_0(\cdot, \cdot) + \langle \cdot, \cdot \rangle_V. \tag{4.11}$$

Under the hypotheses in § 2 on d_0, $\langle \cdot, \cdot \rangle_\lambda$ is norm-equivalent to $\langle \cdot, \cdot \rangle_V$. For $n \geqq 1$, let $P_n(\lambda)$ be the projection of V onto V_n in the inner product $\langle \cdot, \cdot \rangle_\lambda$. Now let $h_1, h_2 \in H$ and note that

$$(\lambda - A)\begin{bmatrix} v^1 \\ v^2 \end{bmatrix} = \begin{bmatrix} h^1 \\ h^2 \end{bmatrix} \tag{4.12}$$

is equivalent to

$$\begin{bmatrix} v^1 \\ v^2 \end{bmatrix} = A^{-1}\left[\lambda \begin{bmatrix} v^1 \\ v^2 \end{bmatrix} - \begin{bmatrix} h^1 \\ h^2 \end{bmatrix}\right]. \tag{4.13}$$

With A^{-1} from (2.10), (4.13) is equivalent to

$$(I + \lambda D_V + \lambda^2 A_0^{-1})v^1 = (\lambda A_0^{-1} + D_V)h^1 + A_0^{-1}h^2 \tag{4.14}$$

and

$$v_2 = \lambda v^1 - h^1. \tag{4.15}$$

If

$$v_n^1 = P_n(\lambda)v^1 \quad \text{and} \quad v_n^2 = P_n(\lambda)v^2, \tag{4.16}$$

it follows from (4.11) and (4.14) that

$$\begin{aligned} \langle e_i, v_n^1 \rangle_\lambda &= \lambda^2 \langle e_i, A_0 A_0^{-1} v^1 \rangle_H + \lambda \langle e_i, D_V v^1 \rangle_V + \langle e_i, v^1 \rangle_V \\ &= \langle e_i, (\lambda A_0^{-1} + D_V)h^1 + A_0^{-1}h^2 \rangle_V, \end{aligned} \tag{4.17}$$

and from (4.15) that

$$\langle e_i, v_n^2 \rangle_\lambda = \langle e_i, v^2 \rangle_\lambda = \lambda \langle e_i, v^1 \rangle_\lambda - \langle e_i, h^1 \rangle_\lambda. \tag{4.18}$$

Now, for $h^1 = h_n^1 \in V_n$, $h^2 = h_n^2 \in V_n$, and v_n^1, v_n^2, h_n^1 and h_n^2 written as in (4.10), (4.17) and (4.18) yield (4.8) and (4.9) again.

This shows that

$$\begin{bmatrix} P_n(\lambda) & 0 \\ 0 & P_n(\lambda) \end{bmatrix} (\lambda - A)^{-1}|_{E_n} = (\lambda - A_n)^{-1}, \tag{4.19}$$

which yields

$$\begin{bmatrix} P_n(\lambda) & 0 \\ 0 & P_n(\lambda) \end{bmatrix} (\lambda - A)^{-1} P_{En} = (\lambda - A_n)^{-1} P_{En} \tag{4.20}$$

where P_{En} *is the E-projection of E onto* E_n. The projection P_{En} can be written

$$P_{En} = \begin{bmatrix} P_{Vn} & 0 \\ 0 & P_{Hn} \end{bmatrix} \tag{4.21}$$

where P_{Vn} is the V-projection onto V_n, as before, and P_{Hn} is the H-projection onto V_n. Since the V-norm is stronger than the H-norm and the norm induced by the λ-inner product is equivalent to the V-norm, it follows from Hypothesis 4.1 that $(\lambda - A_n)^{-1} P_{En}$ converges E-strongly to $(\lambda - A)^{-1}$ as $n \to \infty$. Now, with A_n extended to $E_n^\perp$ as, say $n(P_{En} - I)$, the Trotter–Kato Theorem [K1, Thm. 2.16, p. 504] yields the following.

THEOREM 4.2. *For A_0 coercive, let $T_n(\cdot)$ be the (contraction) semigroup generated on E_n by A_n. Then, for each $t \geqq 0$, $T_n(t)P_{En}$ converges strongly to $T(t)$, uniformly in t for t in bounded intervals.*

In the general case, when A_0 is not coercive, the open-loop generator A is obtained from the dissipative $\tilde{A}$ by the bounded perturbation in (2.12) so that [G3, Thm. 6.6] yields the following generalization of Theorem 4.2.

COROLLARY 4.3. *Let $T_n(\cdot)$ be the semigroup generated on E_n by A_n. Then, for each $t \geqq 0$, $T_n(t)P_{En}$ converges strongly to $T(t)$, uniformly in t for t in bounded intervals.*

THEOREM 4.4. *When A has compact resolvent, $(\lambda - A_n)^{-1}P_{En}$ converges in $L(E)$ to $(\lambda - A)^{-1}$.*

Proof. This follows from (4.20) and a standard result that the projections of a compact linear operator onto a sequence of subspaces converge in norm if the projections converge strongly to the identity, as do P_{En} and $P_n(\lambda)$. □

That the adjoint semigroups also converge strongly follows from an argument entirely analogous to the proof of Theorem 4.2. In particular, equations such as (4.11)-(4.17) are used to show that

$$\begin{bmatrix} P_n(\lambda) & 0 \\ 0 & P_n(\lambda) \end{bmatrix} (\lambda - A^*)^{-1}P_{En} = (\lambda - A_n^*)^{-1}P_{En}. \tag{4.22}$$

In showing this, A^{-*} is used as A^{-1} was used previously. Also, care must be taken to calculate A_n^* with respect to the E-inner product. The results is Theorem 4.5.

THEOREM 4.5. *Let $T_n(\cdot)$ be the sequence of semigroups in Corollary 4.3. Then, for each $t \geqq 0$, $T_n^*(t)P_{En}$ converges strongly to $T^*(t)$, uniformly in t for t in bounded intervals.*

Finally, for the approximation to the actuator influence operator $B \in L(R^m, E)$, recall $B_n \in L(R^m, E_n)$, the operator whose matrix representation is the matrix B^n in (4.5). From (4.3), it follows that

$$B_n = P_{En}B. \tag{4.23}$$

Since B has finite rank m, B_n and B_n^* converge in norm to B and B^*, respectively.

4.2. The approximating optimal control problems. The *nth optimal control problem* is as follows. Given $z_n(0) = (x_n(0), \dot{x}_n(0)) \in E_n$, choose $u \in L_2(0, \infty; R^m)$ to minimize

$$J_n(z_n(0), u) = \int_0^\infty (\langle Q_n z_n(t), z_n(t)\rangle_E + \langle Ru(t), u(t)\rangle_{R^m})\, dt, \tag{4.24}$$

where $Q_n = P_{En}Q|_{En}$. We assume the following hypothesis.

Hypothesis 4.6. For each $n \geqq 1$ and $z_n(0) \in E_n$, there exists an admissible control (Definition 3.1) for (4.6) and (4.24).

A sufficient condition for Hypothesis 4.6 is that, for each n, the system (A_n, B_n) be stabilizable.

By Theorem 3.3, the optimal control $u_n(t)$ has the feedback form

$$u_n(t) = -R^{-1}B_n^*\Pi_n z_n(t) \tag{4.25}$$

where Π_n is a linear operator on E_n, Π_n is nonnegative and self-adjoint, and Π_n satisfies the Riccati equation

$$A_n^*\Pi_n + \Pi_n A_n - \Pi_n B_n R^{-1}B_n^*\Pi_n + Q_n = 0. \tag{4.26}$$

As a result of Hypothesis 4.6, (4.26) has at least one nonnegative, self-adjoint solution. The minimal such solution is the correct Π_n here. If the system (A_n, Q_n) is observable, then Π_n is the unique nonnegative, self-adjoint solution to (4.26) and is positive definite. If we write Π_n as

$$\Pi_n = \begin{bmatrix} \Pi_{0n} & \Pi_{1n} \\ \Pi_{1n}^* & \Pi_{2n} \end{bmatrix}, \tag{4.27}$$

then (4.25) becomes

$$u_n(t) = -R^{-1}B_0^*[\Pi_{1n}^* x_n(t) + \Pi_{2n}\dot{x}_n(t)]. \tag{4.28}$$

The feedback law (4.28) can be written in functional-feedback form, just as in § 3. We have

$$u_n(t) = [u_{1n}(t)u_{2n}(t) \cdots u_{mn}(t)]^T \tag{4.29}$$

where

$$u_{in}(t) = -\langle f_{in}, x_n(t)\rangle_V - \langle g_{in}, \dot{x}_n(t)\rangle_H, \qquad 1 \leqq i \leqq m, \tag{4.30}$$

and

$$f_{in} = \sum_{j=1}^{m} (R^{-1})_{ij}\Pi_{1n}P_{Hn}b_j, \qquad 1 \leqq i \leqq m, \tag{4.31a}$$

$$g_{in} = \sum_{j=1}^{m} (R^{-1})_{ij}\Pi_{2n}P_{Hn}b_j, \qquad 1 \leqq i \leqq m. \tag{4.31b}$$

Of course, f_{in} and g_{in} are the nth approximations to the functional gains f_i and g_i in (3.30).

For numerical solution of the nth problem, we need the matrix representations of these equations. We begin with the following Grammian matrices:

$$\tilde{K}^n = [\langle e_i, e_j\rangle_V] = K^n + [\langle A_1 e_i, e_j\rangle_H] \tag{4.32}$$

and

$$W^n = \begin{bmatrix} \tilde{K}^n & 0 \\ 0 & M^n \end{bmatrix}. \tag{4.33}$$

(*Note.* The matrix W^{-n} will be the inverse of W^n.)

Now recall $Q_n = P_{En}Q|_{En}$. Since $Q = Q^* \in L(E)$ and $E = V \times H$, we can write

$$Q = \begin{bmatrix} Q_0 & Q_1 \\ Q_1^* & Q_2 \end{bmatrix} \tag{4.34}$$

where $Q_0 = Q_0^* \in L(V)$, $Q_1 \in L(H, V)$, and $Q_2 = Q_2^* \in L(H)$. Straightforward calculation shows that

$$Q^n = W^{-n}\tilde{Q}^n \tag{4.35}$$

where Q^n is the matrix representation of Q_n and $\tilde{Q}^n$ is the nonnegative, symmetric matrix

$$\tilde{Q}^n = \begin{bmatrix} \tilde{Q}_0^n & \tilde{Q}_1^n \\ \tilde{Q}_1^{nT} & \tilde{Q}_2^n \end{bmatrix} \tag{4.36}$$

with

$$\tilde{Q}_0^n = [\langle e_i, Q_0 e_j\rangle_V], \quad \tilde{Q}_1^n = [\langle e_i, Q_1 e_j\rangle_V], \quad \tilde{Q}_2^n = [\langle e_i, Q_2 e_j\rangle_H]. \tag{4.37}$$

Also, recall that A_n and B_n are the operators whose matrix representations are given by (4.5), and note that the matrix representations of A_n^* and B_n^* are $W^{-n}(A^n)^T W^n$ and $(B^n)^T W^n$, respectively.

With the matrix representation of Π_n denoted by Π^n, the Riccati operator equation (4.26) is equivalent to the Riccati matrix equation

$$W^{-n}(A^n)^T W^n \Pi^n + \Pi^n A^n - \Pi^n B^n R^{-1}(B^n)^T W^n \Pi^n + Q^n = 0. \tag{4.38}$$

While Π_n is self-adjoint, Π^n in general is not symmetric, but the matrix

$$\tilde{\Pi}^n = W^n \Pi^n \tag{4.39}$$

is symmetric and nonnegative, and positive definite if Π_n is. Premultiplying (4.38) by W^n, we obtain

$$(A^n)^T \tilde{\Pi}^n + \tilde{\Pi}^n A^n - \tilde{\Pi}^n B^n R^{-1}(B^n)^T \tilde{\Pi}^n + \tilde{Q}^n = 0, \tag{4.40}$$

which is the Riccati matrix equation to be solved numerically.

Now we need one more set of matrix equations for the numerical solution of the nth optimal control problem. Since the functional gains f_{in} and g_{in} are elements of V_n, they can be written as

$$f_{in} = \sum_{j=1}^{n} \beta_j^{f_i} e_j \quad \text{and} \quad g_{in} = \sum_{j=1}^{n} \beta_j^{g_i} e_j, \quad i = 1, \cdots, m \tag{4.41}$$

where $\beta^{f_i}, \beta^{g_i} \in R^n$. Partitioning Π^n as in (4.27) and working out the matrix representation of (4.31), we obtain

$$\begin{bmatrix} \beta^{f_1} & \beta^{f_2} & \cdots & \beta^{f_m} \\ \beta^{g_1} & \beta^{g_2} & \cdots & \beta^{g_m} \end{bmatrix} = W^{-n} \tilde{\Pi}^n B^n R^{-1}. \tag{4.42}$$

See [G6] for a derivation of (4.42) that is more natural in terms of finite-dimensional control theory.

5. Convergence. As in § 3, § 5.1 will state some results for the optimal linear regulator problem involving generic linear operators A, B, Q, etc., on an arbitrary real Hilbert space E, and § 5.2 will expand on these results for the particular class of control problems treated in this paper. The proofs of the results in this section are given in [G6].

5.1. Generic approximation results. Let the Hilbert space E and the linear operators A, $T(\cdot)$, B, Q, and R be as in § 3. Suppose that there is a sequence of finite-dimensional subspaces E_n, with the projection of E onto E_n denoted by P_{En}, such that P_{En} converges strongly to the identity as $n \to \infty$, and suppose that there exist sequences of operators $A_n \in L(E_n)$, $B_n \in L(R^m, E_n)$, $Q_n = Q_n^* \in L(E_n)$, $Q_n \geqq 0$, such that we have the following strong convergence. For all $z \in E$ and $t \geqq 0$,

$$\exp(A_n t) P_{En} z \to T(t) z \quad \text{and} \quad \exp(A_n^* t) P_{En} z \to T^*(t) z \tag{5.1}$$

as $n \to \infty$, uniformly in t for t in bounded intervals; for each $u \in R^m$,

$$B_n u \to Bu; \tag{5.2}$$

for each $z \in E$,

$$Q_n P_{En} z \to Qz. \tag{5.3}$$

THEOREM 5.1. *Suppose that for each n there is a nonnegative, self-adjoint linear operator* Π_n *on* E_n *that satisfies the Riccati algebraic equation*

$$A_n^* \Pi_n + \Pi_n A_n - \Pi_n B_n R^{-1} B_n^* \Pi_n + Q_n = 0. \tag{5.4}$$

If there exist positive constants M and β, *independent of n, such that*

$$\|\exp([A_n - B_n R^{-1} B_n^* \Pi_n] t)\| \leqq M e^{-\beta t}, \qquad t \geqq 0, \tag{5.5}$$

and if $\|\Pi_n\|$ *is bounded uniformly in n, then the Riccati algebraic equation* (3.3) *has a nonnegative self-adjoint solution* Π, *and, for each* $z \in E$,

$$\Pi_n P_{En} z \to \Pi z \tag{5.6}$$

and

$$\exp([A_n - B_n R^{-1} B_n^* \Pi_n]t) P_{En} z \to S(t)z \tag{5.7}$$

uniformly in $t \geqq 0$, *where* $S(\cdot)$ *is the semigroup generated by* $A - BR^{-1}B^*\Pi$. *If there exists a positive constant* δ, *independent of n, such that*

$$Q_n \geqq \delta, \tag{5.8}$$

then $\|\Pi_n\|$ *being bounded uniformly in n guarantees the existence of positive constants M and* β *for which* (5.5) *holds for all n.*

THEOREM 5.2. *The strong convergence in* (5.6) *implies uniform norm convergence of the optimal feedback laws*:

$$\|B_n^* \Pi_n P_{En} - B^* \Pi\| \to 0 \quad as\ n \to \infty. \tag{5.9}$$

THEOREM 5.3. *Assume the hypotheses of Theorem* 5.1 *but do not assume* (5.5) *or* (5.8). *If* $\|\Pi_n\|$ *is bounded uniformly in n, then the Riccati algebraic equation* (3.3) *has a nonnegative self-adjoint solution* Π *and, for each* $z \in E$, $\Pi_n P_{En} z$ *converges weakly to* Πz.

5.2. Convergence of the approximating optimal control problems of § 4.2. For the rest of this section, A_0, A_1, A, $T(t)$, B_0, and B are the operators defined in § 2. The operators A_n, B_n, Q_n, and Π_n are the operators in the approximation scheme of § 4. Also, $\Pi \in L(E_n, E_n)$ is the minimal nonnegative, self-adjoint solution of the Riccati operator equation (4.26). According to Corollary 4.3 and Theorem 4.5, the Ritz-Galerkin approximation scheme presented in § 4.1 converges as required in (5.1); (5.2) and (5.3) follow from (4.23) and the definition $Q_n = P_{En} Q|_{En}$ in § 4.2. Also, Hypothesis 4.6 guarantees for each n the existence of the required solution of the Riccati equation (5.4) in Theorem 5.1.

Since Π_n is nonnegative and self-adjoint, its eigenvalues, which are also the eigenvalues of its matrix representation, are real and nonnegative, and its norm is equal to its maximum eigenvalue.

THEOREM 5.4. *If Q is E-coercive and* $d_0 = 0$ (*i.e., there is no open-loop damping*), *then there is no nonnegative self-adjoint solution of the Riccati operator equation* (3.3), *and*

$$\|\Pi_n\| \to \infty \quad as\ n \to \infty. \tag{5.10}$$

THEOREM 5.5. *Suppose that* A_0 *and* $d_0(\cdot, \cdot)$ *are both H-coercive. Then there exist positive constants* M_1 *and* α_1, *independent of n, such that*

$$\|\exp[A_n t]\| \leqq M_1 e^{-\alpha_1 t}, \qquad t \geqq 0. \tag{5.11}$$

THEOREM 5.6. *Suppose that* A_0 *has an invariant subspace* V_0 *that is also invariant under the damping map* D_V, *that* $E_0 = V_0 \times V_0$ *is a stabilizable subspace for the control system, and that the restrictions of* A_0 *and* $d_0(\cdot, \cdot)$ *to* $V_0^\perp$ *are both H-coercive. Also, suppose that* V_0 *has finite dimension* n_0 *and that, for each* $n \geqq n_0$ *in the approximation scheme, the first* n_0 e_i*'s span* V_0 *and the rest are orthogonal to* V_0 *in both V and H.*

(i) *Then* (3.3) *has a nonnegative self-adjoint solution* Π, *and for each* $n \geqq n_0$, (5.4) *has a nonnegative self-adjoint solution* Π_n. *Also*, Π_n *is bounded uniformly in n, so that* Π_n *converges to* Π *weakly, as in Theorem* 5.3.

(ii) *If E_0 and $E_0^\perp$ (the E-orthogonal complement of E_0) are invariant under Q, and if the part of the open-loop system on E_0 is observable with the measurement Qz, then* (5.5)–(5.7) *hold as in Theorem* 5.1.

Remark 5.7. In applications, the subspace V_0 in Theorem 5.6 usually contains rigid-body modes. The theorem includes the case where both A_0 and d_0 are H-coercive on all of V (no rigid-body modes and all modes damped). In this case, V_0 is the trivial subspace.

Remark 5.8. Otherwise, for applications to flexible structures, Theorem 5.6 usually requires two things: first, modal damping must be modeled for the structure, so that the natural modes remain uncoupled in the open-loop system; second, the natural mode shapes must be used for the basis functions in the approximating optimal control problems. Although these requirements may seem restrictive from a mathematical standpoint, such modeling and approximation predominate in engineering practice. Also, we get our strongest convergence results under these conditions. For applications where the basis vectors are not the natural mode shapes, the following theorem is useful.

THEOREM 5.9. *Suppose that $A_0+B_0B_0^*$ and $d_0+B_0B_0^*$ are H-coercive. Then* (3.3) *has a nonnegative solution* Π, *for each n* (5.4) *has a nonnegative self-adjoint solution* Π_n, *and* $\|\Pi_n\|$ *is bounded uniformly in n. Hence, Theorem* 5.3 *applies. Furthermore, if Q is E-coercive, then* (5.5)–(5.8) *hold in Theorem* 5.1.

THEOREM 5.10. *If* (5.6) *holds for each $z \in E$, then*

$$\|f_{in}-f_i\|_V \to 0, \quad \|g_{in}-g_i\|_H \to 0 \quad \text{as } n \to \infty \tag{5.12}$$

where f_i and g_i are the functional gains in (3.32), *and f_{in} and g_{in} are the approximating functional gains in* (4.31) *and* (4.41).

Note that (5.9) and (5.12) are equivalent.

6. Example.

6.1. The control system. One end of the uniform Euler–Bernoulli beam in Fig. 6.1 and Table 6.1 is attached rigidly (cantilevered) to a rigid hub (disc), which is free to rotate about its center, point 0, which is fixed. Also, a point mass m_1 is attached to the other end of the beam. The control is a torque u applied to the disc, and all motion is in the plane.

The angle θ represents the rotation of the disc (the rigid-body mode), $w(t, s)$ is the elastic deflection of the beam from the rigid-body position, and $w_1(t)$ is the displacement of m_1 from the rigid-body position. For technical reasons, we do not yet impose the condition $w_1(t) = w(t, l)$; more will be said on this later.

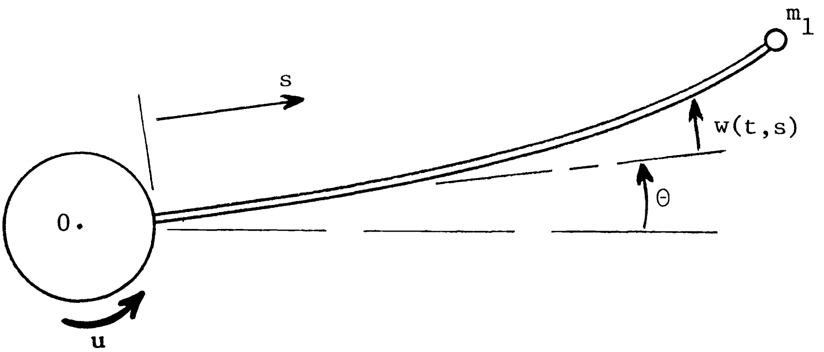


FIG. 6.1. *Control system.*

TABLE 6.1
Structural data.

r = hub radius	10 in
l = beam length	100 in
I_0 = hub moment of inertia about axis perpendicular to page through 0	100 slug in^2
m_b = beam mass per unit length	.01 slug/in
m_1 = tip mass	1 slug
EI = product of elastic modulus and second moment of cross section for beam	13,333 slug in^3/sec^2
fundamental frequency of undamped structure	.9672 rad/sec

The control problem is to stabilize rigid-body motions and linear (small) transverse elastic vibrations about the state $\theta = 0$ and $w = 0$. Our linear model assumes not only that the elastic deflection of the beam is linear but also that the axial inertial force produced by the rigid-body angular velocity has negligible effect on the bending stiffness of the beam. The rigid-body angle need not be small.

For this example, it is a straightforward exercise to derive the three coupled differential equations of motion in θ, w, and w_1, and they do have the form (2.1′). However, to emphasize the fact that we do not use the explicit partial differential equations, we will not write these equations here. Rather, we will write only what is normally needed in applications: the kinetic and strain-energy functionals, the damping functional, and the actuator influence operator.

Remark 2.1 applies to this example, and to most examples with complex structures. The generalized displacement vector is

$$x = (\theta, w, w_1) \in H_0 = R \times L_2(0, l) \times R. \tag{6.1}$$

The kinetic energy in the system is

$$\text{Kinetic Energy} = \tfrac{1}{2}\langle \dot{x}, \dot{x} \rangle_H \tag{6.2}$$

where H is H_0 with the inner product

$$\begin{aligned}\langle x, \hat{x} \rangle_H = m_b \int_0^l [w + (r+s)\theta][\hat{w} + (r+s)\hat{\theta}]\, ds + I_0\theta\hat{\theta} + m_1[w_1 + (r+l)\theta]\\ \cdot [\hat{w}_1 + (r+l)\hat{\theta}].\end{aligned} \tag{6.3}$$

As in most applications, we need not write the mass operator explicitly, but there exists a unique self-adjoint linear operator M_0 on H_0 such that

$$\langle x, \hat{x} \rangle_H = \langle M_0 x, \hat{x} \rangle_{H_0}. \tag{6.4}$$

It is easy to see that M_0 is bounded and coercive. Hence H_0 and H have equivalent norms. The input operator for (2.1′) (which maps R to H_0) is

$$B_0 = (1, 0, 0). \tag{6.5}$$

Remark 2.2 also applies here. The only strain energy is in the beam and is given by

$$\text{Strain Energy} = \tfrac{1}{2}a(x, x) \tag{6.6}$$

with

$$a(x, \hat{x}) = EI \int_0^l w''\hat{w}''\, ds \tag{6.7}$$

where $(\cdot)''=\partial^2(\cdot)/\partial s^2(\cdot)$. To make $a(\cdot,\cdot)$ into an inner product, we must account for rigid-body rotation. Thus we set

$$\langle x,\hat{x}\rangle_V=a(x,\hat{x})+\theta\hat{\theta} \tag{6.8}$$

and define

$$V=\{x=(\theta,\phi,\phi(1)):\ \phi\in H^2(0,l),\ \phi(0)=\phi'(0)=0\}. \tag{6.9}$$

Also, we have

$$\langle x,\hat{x}\rangle_V=a(x,\hat{x})+\langle B_0B_0^*x,\hat{x}\rangle_{H_0} \tag{6.10}$$

so that $A_1=B_0B_0^*$. *But we need neither A_1 nor A_0 explicitly. We need only* (6.7) *and* (6.8), *along with* (6.3), *to compute the required inner products.*

We assume that the beam has Voigt–Kelvin viscoelastic damping [C2], so that the damping operator in (2.1) is

$$D_0=c_0A_0 \tag{6.11}$$

where c_0 is a constant. This means that the damping functional is

$$d_0(x,\hat{x})=c_0a(x,\hat{x}),\qquad x,\hat{x}\in V. \tag{6.12}$$

6.2. The optimal control problem. We take $Q=I$ in the performance index in (3.1). This means that the state weighting term $\langle Qz,z\rangle_E$ is twice the total energy in the structure plus the square of the rigid-body rotation. Since there is one input, the control weighting R is a scalar.

According to (3.31), the optimal control has the feedback form

$$u(t)=-\langle f,x(t)\rangle_V-\langle g,\dot{x}(t)\rangle_H \tag{6.13}$$

where $x(t)$ has the form (6.1), and

$$f=(\alpha_f,\phi_f,\beta_f)=R^{-1}\Pi_1B_0\in V,\qquad g=(\alpha_g,\phi_g,\beta_g)=R^{-1}\Pi_2B_0\in H. \tag{6.14}$$

Note that $\beta_f=\phi_f(l)$ is not used in the control law—recall (6.7) and (6.8).

6.3. Approximation. Our approximation of the distributed model of the structure is based on a finite-element approximation of the beam that uses Hermite cubic splines as basis functions [S1], [S4]. Because the basis vectors e_j in the approximation scheme in § 4 must be in the space V defined in (6.9), we write them as

$$e_1=(1,0,0),\quad e_j=(0,\phi_j,\phi_j(1)),\quad j=2,3,\cdots \tag{6.15}$$

where the ϕ_j's are the cubic splines. When we use n_e elements to approximate the beam, there are $2n_e$ linearly independent splines. Thus, with the rigid-body mode, the order of approximation is $n=2n_e+1$.

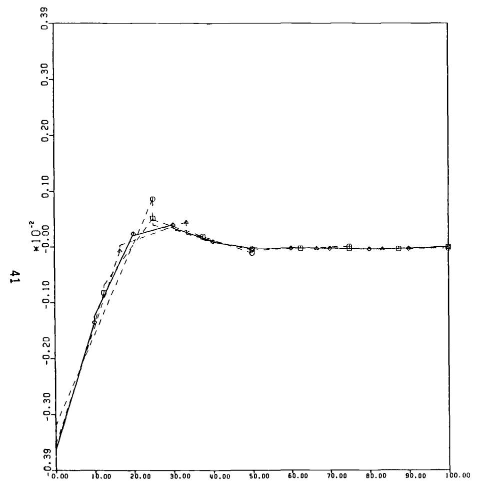
For the numerical solution to the optimal control problem, we have only to plug into the formulas of § 4. The matrices in (4.3) are calculated according to (6.3), (6.7), and (6.8), with B_0 given by (6.5). The matrices A^n and B^n are given by (4.5) and, since $Q=I$, the matrix $\tilde{Q}^n$ is the W^n in (4.33). With these matrices, we solve the Riccati equation (4.40) and use (4.41) and (4.45) to compute the approximations to the functional gains, which are

$$f_n=(\alpha_{fn},\phi_{fn},\beta_{fn}),\qquad g_n=(\alpha_{gn},\phi_{gn},\beta_{gn}). \tag{6.16}$$

For convergence, we satisfy all the hypotheses of Theorem 5.9. In particular, since Q is the identity on E, it is coercive. Theorem 5.9 implies that the solutions to the finite-dimensional Riccati equations converge as in Theorem 5.1 and that the functional control gains converge as in Theorem 5.10.

6.4. Numerical results. Figures 6.2(a) and 6.2(b) show the computed functional gain kernels ϕ''_{fn} and ϕ_{gn} for the damping coefficient $c_0 = 10^{-4}$, the control weighting $R = .05$, and $n_e = 4, 6, 8$, and 10 beam elements. Table 6.2 lists the corresponding scalar components of the gains. We have plotted ϕ''_{fn} because the second derivative appears in the strain-energy inner product in (6.7) and (6.8) and ϕ_{fn} converges in $H^2(0, l)$. Note that, since the Hermite cubic splines have discontinuous second derivatives at the nodes, the approximations to ϕ''_f are discontinuous at the nodes. Although H^2-convergence guarantees only L_2-convergence for ϕ''_{fn}, it can be shown that ϕ''_{fn} converges uniformly on $[0, l]$ for this problem.

Tables 6.2 and 6.3 omit β_{fn} to emphasize the fact that it does not appear in the feedback law and the fact that the convergence of β_{fn} is not an independent piece of information about the convergence of the control gains; since $\phi_{fn}(0) = \phi'_{fn}(0) = 0$, the convergence of ϕ''_{fn} implies the convergence of $\beta_{fn} = \phi_{fn}(l)$. On the other hand, although $\beta_{gn} = \phi_{gn}(l)$ for each n, the H-norm convergence of g_n does not enforce this condition in the limit, as the V-norm convergence of f_n enforces $\beta_f = \phi_f(l)$. Hence, as far as we can tell from our results in §§ 3 and 5, β_{fn} is an independent indicator of the convergence of the control gains, as well as being used in the control law in (6.13). However, the behavior of ϕ_{gn} in Fig. 6.2(b) suggests that g_n converges in V. Stronger results on the continuity of ϕ_g and the convergence of $\phi_{gn}(l)$ should follow from a theorem stating that, because the open-loop semigroup generator A is analytic, the solution to the



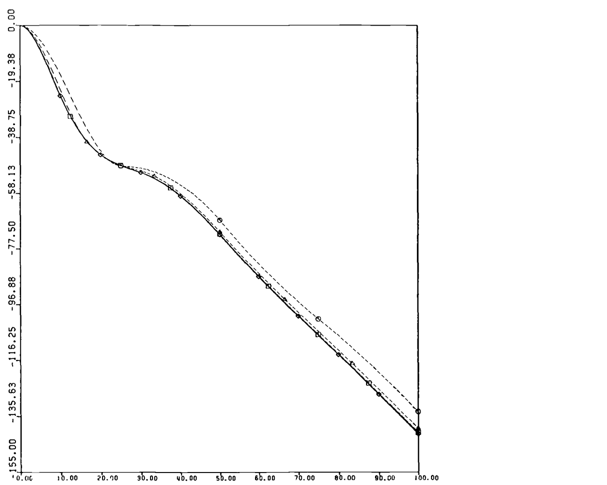


FIG. 6.2(a). *Functional control gain component* ϕ''_{fn}. *Damping coefficient* $c_0 = 10^{-4}$; *control weighting* $R = .05$; *number of elements* $n_e = 4, 6, 8, 10$.

FIG. 6.2(b). *Functional control gain component* ϕ_{gn}. *Damping coefficient* $c_0 = 10^{-4}$; *control weighting* $R = .05$; *number of elements* $n_e = 4, 6, 8, 10$.

TABLE 6.2

Scalar components of functional control gains. Damping coefficient $c_0 = 10^{-4}$; *control weight* $R = .05$.

n_e	α_{fn}	α_{gn}	β_{gn}
4	4.4721	1.2440	−133.87
6	4.4721	1.2973	−139.69
8	4.4721	1.3106	−141.15
10	4.4721	1.3141	−141.54

infinite-dimensional Riccati equation maps all of E into $D(A^*)$. The fact that $\phi''_{fn}(l)$ converges to zero in Fig. 6.2(a) also suggests such a theorem, but we have not proved it.

With the state weighting Q fixed, the two factors that determine the rate of convergence are c_0 and R. Although we have used splines to approximate the beam, the relation between the convergence rate and c_0 and R probably can be interpreted best in terms of the number of natural modes of the structure that the optimal infinite-dimensional controller really controls. Strictly speaking, the controller controls all modes, but the functional gains lie essentially in the span of some finite number of modes. This would be the number of modes required for convergence of the gains if we used the natural modes as the basis vectors in the approximation. The rest of the modes are practically (but not exactly) orthogonal to the functional gains, so that the optimal feedback law essentially ignores them. In general, the lighter the damping, the more modes that will be controlled for given Q and R; the cheaper the control, the more modes that will be controlled for given Q and c_0. The question of the convergence of the finite-element approximation to the functional gains then becomes a question of how many modes the optimal control really wants and how many elements it takes to approximate those modes.

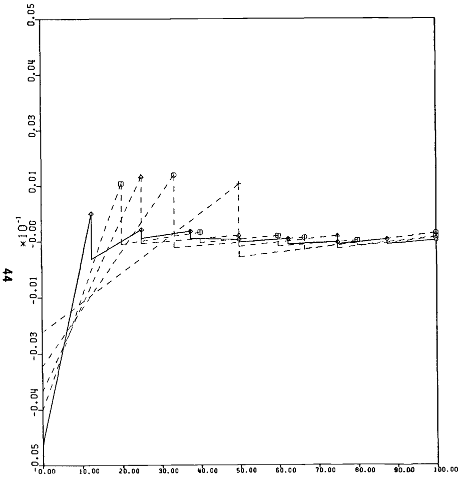
Numerical experience with optimal control of flexible structures has shown this modal interpretation of the convergence of the approximating control laws to be very useful, and that it is difficult to improve on the natural modes as basis vectors for the approximation scheme (see [G5]). However, whether the natural modes are always or almost always the best basis vectors is an open question. We use the cubic splines here to demonstrate that a standard finite-element approximation works quite well. Also, to use the natural modes as basis vectors here, we first would have to compute them using a finite-element approximation—as in most real problems—and we do not know in advance which or how many modes are needed. On the other hand, if the most important natural modes are determined from experiment, then modal approximation should be best.

Figures 6.3(a) and 6.3(b) and Table 6.3 represent attempts to compute an optimal control law for the structure when $R=.05$ but $c_0=0$. Since Q is the identity operator in E and hence coercive, Theorem 5.4 says that no optimal control law exists and that the norm of the solution to the finite-dimensional Riccati equation grows without bound as the number of elements increases. This is reflected in the nonconvergence of α_{gn}, ϕ_{gn}, and β_{gn}, although α_{fn} converges and the convergence of ϕ''_{fn} is unclear.

TABLE 6.3
Scalar components of functional control gains. Zero damping; control weighting $R=.05$.

n_e	α_{fn}	α_{gn}	β_{gn}
2	4.4721	1.0516	−112.23
3	4.4721	1.3061	−140.18
4	4.4721	1.4758	−159.11
5	4.4721	1.5996	−172.64
8	4.4721	1.8407	−199.39

In applications where the structural damping is not known, except that it is very light, it is tempting and not uncommon engineering practice to assume zero damping in the design of a control law for the first few modes, while trusting whatever damping is in the higher modes to take care of them. However, if high performance requirements



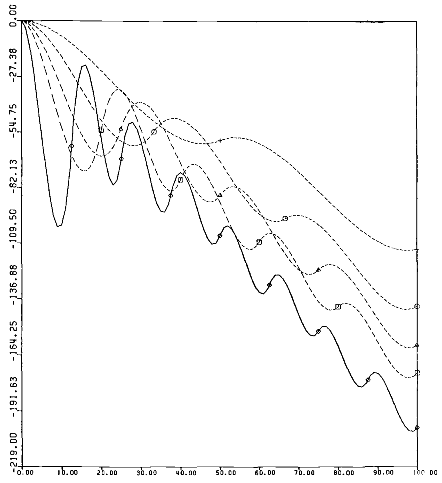


FIG. 6.3(a). *Functional control gain component* ϕ''_{fn}. *Zero damping*; *control weighting* $R = .05$; *number of elements* $n_e = 2, 3, 4, 5, 8$.

FIG. 6.3(b). *Functional control gain component* ϕ_{gn}. *Zero damping*; *control weighting* $R = .05$; *number of elements* $n_e = 2, 3, 4, 5, 8$.

(large Q) or coupling between modes in the closed-loop system necessitate a control law based on a more accurate approximation of the structure, Theorem 5.4 and the current example warn that the higher-order control laws are likely to be meaningless and rather strange if no damping is modeled.

We should note that we have seen similar problems [G10] where Π_n remains bounded and the gains converge for zero damping but finite-rank Q. In such cases, Theorem 5.3 says that an optimal control law exists for the distributed model of the structure and that the finite-dimensional control laws converge to an optimal infinite-dimensional control law. Also, Balakrishnan [B2] has shown that an infinite-dimensional optimal control law exists for no damping when $Q = BB^*$.

7. The optimal infinite-dimensional estimator, compensator, and closed-loop system.

7.1. The generic problem. Let A, $T(t)$, and B be as in § 3.1, with E an arbitrary real Hilbert space. The differential equation corresponding to (3.2) is, of course,

$$\dot{z}(t) = Az(t) + Bu(t), \qquad t > 0. \tag{7.1}$$

We assume that we have a p-dimensional measurement vector $y(t)$ given by

$$y(t) = C_0 u(t) + Cz(t) \tag{7.2}$$

where $C_0 \in L(R^m, R^p)$ and $C \in L(E, R^p)$ for some positive integer p.

DEFINITION 7.1. For any $\hat{F} \in L(R^p, E)$, the system

$$\dot{\hat{z}}(t) = A\hat{z}(t) + Bu(t) + \hat{F}[y(t) - C_0 u(t) - C\hat{z}(t)], \qquad t > 0, \tag{7.3}$$

will be called an *observer, or estimator* (we use the terms interchangeably) for the system (7.1)-(7.2). Let $\hat{S}(t)$ be the semigroup generated by $A - \hat{F}C$. The observer in (7.3) is strongly (uniformly exponentially) stable if $\hat{S}(t)$ is strongly (uniformly exponentially) stable.

To justify this definition, we write

$$e(t) = z(t) - \hat{z}(t) \tag{7.4}$$

and, with (7.1)-(7.3), obtain

$$e(t)=\hat{S}(t)e(0), \qquad t \geqq 0. \tag{7.5}$$

Of course, an observer, or estimator, is necessary because the full state $z(t)$ will not be available for direct feedback, and the feedback control must be based on an estimate of $z(t)$. When, as in this paper, the desired control law has the form

$$u(t)=-Fz(t) \tag{7.6}$$

for some $F \in L(E, R^m)$, the observer in (7.3) can be used to construct $\hat{z}(t)$ from the measurement in (7.2) and then the control law in (7.6) can be applied to $\hat{z}(t)$. The control applied to the system is then

$$u(t)=-F\hat{z}(t), \tag{7.7}$$

and the resulting closed-loop system is

$$\begin{bmatrix} z(t) \\ \hat{z}(t) \end{bmatrix} = S_{\infty\infty}(t) \begin{bmatrix} z(0) \\ \hat{z}(0) \end{bmatrix}, \qquad t \geqq 0 \tag{7.8}$$

where $S_{\infty\infty}(t)$ is the semigroup generated on $E \times E$ by the operator

$$A_{\infty\infty}=\begin{bmatrix} A & -BF \\ \hat{F}C & [A-BF-\hat{F}C] \end{bmatrix}, \qquad D(A_{\infty\infty})=D(A)\times D(A). \tag{7.9}$$

With the estimator error $e(t)$ defined by (7.4), it is easy to show that (7.8) is equivalent to (7.5) and

$$\dot{z}(t)=(A-BF)z(t)+BFe(t), \qquad t>0 \tag{7.10}$$

where $(A-BF)$ generates a semigroup $S(t)$ on E. Also, it is easy to prove the following.

THEOREM 7.2. *Suppose that there exist positive constants M_1, M_2, α_1, and α_2 such that*

$$\|S(t)\| \leqq M_1 e^{-\alpha_1 t}, \quad \|\hat{S}(t)\| \leqq M_2 e^{-\alpha_2 t}, \quad t \geqq 0. \tag{7.11}$$

Then, for each real $\alpha_3 < \min\{\alpha_1, \alpha_2\}$, there exists a constant M_3 such that

$$\|S_{\infty\infty}(t)\| \leqq M_3 e^{-\alpha_3 t}, \qquad t \geqq 0. \tag{7.12}$$

Also,

$$\sigma(A_{\infty\infty})=\sigma(A-BF)\cup\sigma(A-F\hat{C}) \tag{7.13}$$

where $\sigma(A_{\infty\infty})$ is the spectrum of $A_{\infty\infty}$.

The observer in (7.3) and the control law in (7.7) constitute a compensator for the control system in (7.1) and (7.2). The transfer function of this compensator is

$$\Phi(s)=-F(sI-[A-BF+\hat{F}(C_0F-C)])^{-1}\hat{F}, \tag{7.14}$$

which is an $m \times p$ matrix function of the complex variable s. When E has infinite dimension, the compensator transfer function is irrational, except in degenerate, usually unimportant cases.

The foregoing definitions of this section and Theorem 7.2 are straightforward generalizations to infinite dimensions of observer-controller results in finite dimensions. Balas [B3] and Schumacher [S2] have used similar extensions.

Now suppose that $\hat{F}$ is chosen as

$$\hat{F}=\hat{\Pi}C^*\hat{R}^{-1} \tag{7.15}$$

where $\hat{\Pi} \in L(E, E)$ is the minimal nonnegative self-adjoint solution to the Riccati equation

$$A\hat{\Pi} + \hat{\Pi}A^* - \hat{\Pi}C^*\hat{R}^{-1}C\hat{\Pi} + \hat{Q} = 0, \tag{7.16}$$

with $\hat{Q} \in L(E, E)$ nonnegative and self-adjoint and $\hat{R} \in L(R^p, R^p)$ symmetric and positive definite. Theorem 3.3 (with A, B, Q, R, Π, and $S(t)$ replaced by A^*, C^*, $\hat{Q}$, $\hat{R}$, $\hat{\Pi}$, and $\hat{S}^*(t)$) gives sufficient conditions for $\hat{\Pi}$ to exist and for the semigroup $\hat{S}^*(t)$—and equivalently its adjoint, the $\hat{S}(t)$ generated by $A - \hat{\Pi}C^*\hat{R}^{-1}C$—to be uniformly exponentially stable.

DEFINITION 7.3. When the control gain operator is

$$F = R^{-1}B^*\Pi, \tag{7.17}$$

with Π the solution to the Riccati equation (3.3) and the observer gain operator is given by (7.15) and (7.16), we will call the compensator consisting of the observer in (7.3) and the control law in (7.7) the *optimal infinite-dimensional compensator*, and (7.8) *the optimal closed-loop system.* (See Fig. 7.1.)

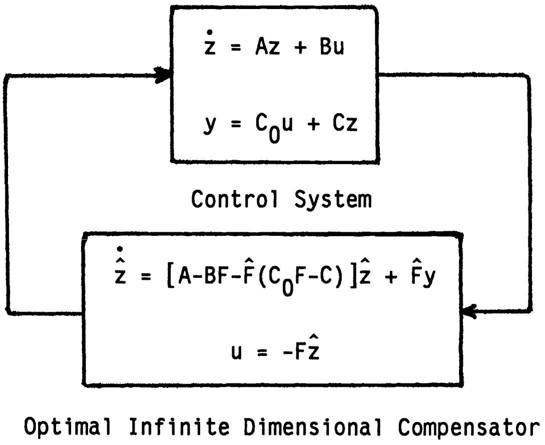

$\dot{z} = Az + Bu$
$y = C_0u + Cz$
Control System
$\dot{\hat{z}} = [A-BF-\hat{F}(C_0F-C)]\hat{z} + \hat{F}y$
$u = -F\hat{z}$
Optimal Infinite Dimensional Compensator

FIG. 7.1. *Optimal closed-loop system.*

Remark 7.4. The infinite-dimensional observer defined by (7.3), (7.15), and (7.16) is the optimal estimator for the stochastic version of (7.1) and (7.2) when (7.1) is disturbed by a stationary Gaussian white noise process with zero mean and covariance operator $\hat{Q}$ and the measurement in (7.2) is contaminated by similar noise with covariance $\hat{R}$. For infinite-dimensional stochastic estimation and control, see [B1], [C4]. When the state weighting operator Q in (3.1) is trace class, the optimal infinite-dimensional compensator minimizes the time-average of the expected steady-state value of the integrand in (3.1). Existing theory for stochastic control of infinite-dimensional systems requires trace-class Q, but we have a well-defined compensator for any bounded nonnegative self-adjoint Q and $\hat{Q}$, as long as the solutions to the Riccati equations exist. As the next two sections show (without assuming trace-class $\hat{Q}$), the infinite-dimensional compensator is the limit of a sequence of finite-dimensional compensators, each of which can be interpreted as an optimal LQG compensator for a finite-dimensional model of the structure. Therefore, we do not require trace-class Q in our definition of the optimal compensator, even though this compensator solves a precise optimization problem only when Q is trace class.

This paper is concerned primarily with how the finite-dimensional compensators converge to the infinite-dimensional compensator, and the analysis of this convergence requires only the theory of infinite-dimensional Riccati equations for deterministic optimal control problems and the corresponding approximation theory. While the stochastic interpretation of the infinite-dimensional compensator and, in § 8.2, of the finite-dimensional estimators should be motivational, nothing in the rest of the paper depends on a stochastic formulation. We assume that the operators Q, R, $\hat{Q}$, and $\hat{R}$ are determined by some design criteria. In many engineering applications, deterministic criteria such as the stability margin and robustness of the closed-loop system, rather than a stochastic performance index and an assumed noise model, govern the choice of Q, R, $\hat{Q}$, and $\hat{R}$.

7.2. Application to structures. For the rest of the paper, $E = V \times H$ as in § 2, and A and B are the operators defined there.

The measurement operator C in (7.2) now must have the form

$$C = [C_1 \quad C_2] \tag{7.18}$$

where $C_1 \in L(V, R^p)$ and $C_2 \in L(H, R^p)$. Hence, if we denote by $(C(x, \dot{x}))_i$ the ith component to the p-vector $C(x, \dot{x})$, for $(x, \dot{x}) \in E$, then there must exist $c_{1i} \in V$ and $c_{2i} \in H$ such that

$$(C(x, \dot{x}))_i = \langle C_{1i}, x \rangle_V + \langle C_{2i}, \dot{x} \rangle_H, \qquad i = 1, \cdots, p. \tag{7.19}$$

Also, the estimator gain operator F is given by

$$\hat{F}y = \sum_{i=1}^{p} (\hat{f}_i, \hat{g}_i) y_i \tag{7.20}$$

for $y = [y_1 y_2 \cdots y_p]^T \in R^p$, where the *functional estimator gains* $\hat{f}_i$ and $\hat{g}_i$ are elements of V and H, respectively.

For the optimal estimator gains, we can partition $\hat{\Pi}$ as

$$\hat{\Pi} = \begin{bmatrix} \hat{\Pi}_0 & \hat{\Pi}_1 \\ \hat{\Pi}_1^* & \hat{\Pi}_2 \end{bmatrix} \tag{7.21}$$

and use (7.15) and (7.19) to get

$$\hat{f}_i = \sum_{j=1}^{p} (\hat{R})_{ij} (\hat{\Pi}_0 c_{1j} + \hat{\Pi}_1 c_{2j}), \tag{7.22a}$$

$$\hat{g}_i = \sum_{j=1}^{p} (\hat{R}^{-1})_{ij} (\hat{\Pi}_1^* c_{1j} + \hat{\Pi}_2 c_{2j}), \qquad i = 1, 2, \cdots, p. \tag{7.22b}$$

Now let us partition $\hat{Q}$ as in (4.34):

$$\hat{Q} = \begin{bmatrix} \hat{Q}_0 & \hat{Q}_1 \\ \hat{Q}_1^* & \hat{Q}_2 \end{bmatrix}. \tag{7.23}$$

In the optimal control problem, we almost always have a nonzero Q_0 because this operator penalizes the generalized displacement. For the results in this paper, $\hat{Q}_0$ can be nonzero in the observer problem, and, as in the control problem, some of the strongest convergence results for finite-dimensional approximations can be proved only for coercive $\hat{Q}$. However, if the observer is to be thought of as an optimal filter, then $\hat{Q}$ should be the covariance operator of the noise that disturbs (2.1). In this case, $\hat{Q}_0 = 0$ and $\hat{Q}_1 = 0$.

8. Approximation of the infinite-dimensional estimator.

8.1. The approximating finite-dimensional estimators. Here, the scheme for the approximation of the flexible structure is that in § 4.

Hypothesis 8.1. There exists a sequence $C_n \in L(E_n, R^p)$ such that

$$\|C_n P_{En} - C\| \to 0 \quad \text{as } n \to \infty \tag{8.1}$$

and a sequence $\hat{Q}_n \in L(E_n)$, $\hat{Q}_n^* = \hat{Q}_n \geqq 0$, such that

$$\hat{Q}_n P_{En} \to \hat{Q} \quad \text{strongly as } n \to \infty. \tag{8.2}$$

Hypothesis 8.2. For each n, the system (A_n^*, C_n^*) is stabilizable. In particular, any unstable modes of the system (C_n, A_n) are observable.

The nth *observer*, or nth *estimator*, is

$$\dot{\hat{z}}_n = A_n \hat{z}_n + B_n u + \hat{F}_n(y - C_0 u - C_n \hat{z}_n) \tag{8.3}$$

where the estimator gain $\hat{F}_n$ is

$$\hat{F}_n = \hat{\Pi}_n C_n^* \hat{R}^{-1}, \tag{8.4}$$

and $\hat{\Pi}_n$ is the nonnegative self-adjoint solution to the Riccati operator equation

$$A_n \hat{\Pi}_n + \hat{\Pi}_n A_n^* - \hat{\Pi}_n C_n^* \hat{R}^{-1} C_n \hat{\Pi}_n + \hat{Q}_n = 0. \tag{8.5}$$

Hypothesis 8.2 implies that such a solution exists and is unique.

On-line computations will be based on the equivalent differential equation

$$\dot{\hat{\eta}} = A^n \hat{\eta} + B^n u + \hat{F}^n(y - C_0 u - C^n \hat{\eta}) \tag{8.6}$$

where $\hat{\eta}(t) \in R^{2n}$, A^n and B^n are the matrix representations of the operators A_n and B_n, as in § 4, and C^n is the matrix representation of C_n. The $2n \times p$ gain matrix $\hat{F}^n$ is

$$\hat{F}^n = \hat{\Pi}^n W^{-n} (C^n)^T \hat{R}^{-1} \tag{8.7}$$

where W^n is the $2n \times 2n$ Grammian matrix in (4.33) and $\hat{\Pi}^n$ satisfies

$$A^n \hat{\Pi}^n + \hat{\Pi}^n W^{-n} (A^n)^T W^n + \hat{\Pi}^n W^{-n} (C^n)^T \hat{R}^{-1} C^n \hat{\Pi}^n + Q^n = 0, \tag{8.8}$$

with $\hat{Q}^n$ the matrix representation of $\hat{Q}_n$. The relationship between $\hat{z}_n = (\hat{x}_n, \dot{\hat{x}}_n)$ and $\hat{\eta}$ is, of course,

$$\hat{x}_n(t) = \sum_{i=1}^{n} \hat{\xi}_i(t) e_i \tag{8.9}$$

and

$$\hat{\eta} = [\hat{\xi}^T \dot{\hat{\xi}}^T]^T. \tag{8.10}$$

It is straightforward to show that (8.7) is the matrix representation of (8.4) and that the $2n \times 2n$ Riccati matrix equation (8.8) is the matrix representation of (8.5), with $\hat{\Pi}^n$ the matrix representation of $\hat{\Pi}_n$.

In (4.39), we defined the symmetric matrix $\tilde{\Pi}^n = W^n \Pi^n$ and then obtained the Riccati equation (4.40) to solve for $\tilde{\Pi}^n$. We proceed in a similar fashion here, but with an interesting difference. Since $\hat{\Pi}_n$ and $\hat{Q}_n$ are nonnegative self-adjoint operators on E_n and $\hat{\Pi}^n$ and Q^n are their matrix representations, the matrices $W^n \hat{\Pi}^n$ and $W^n \hat{Q}^n$ are nonnegative and symmetric. Hence, the matrices

$$\tilde{\hat{\Pi}}^n = \hat{\Pi}^n W^{-n} \tag{8.11}$$

and

$$\tilde{\hat{Q}}^n = \hat{Q}^n W^{-n} \tag{8.12}$$

are nonnegative and symmetric. Substitution of (8.11) and (8.12) into (8.7) and (8.8) yields

$$\hat{F}^n = \tilde{\hat{\Pi}}^n (C^n)^T \hat{R}^{-1} \tag{8.13}$$

and

$$A^n \tilde{\hat{\Pi}}^n + \tilde{\hat{\Pi}}^n (A^n)^T - \tilde{\hat{\Pi}}^n (C^n)^T \hat{R}^{-1} C^n \tilde{\hat{\Pi}}^n + \tilde{\hat{Q}}^n = 0, \tag{8.14}$$

the Riccati matrix equation to be solved numerically in the nth approximation to the infinite-dimensional estimator. In view of the relationship between (8.5) and (8.8) and the relationship between (8.8) and (8.14), we see that Hypothesis 8.1 guarantees the existence of a unique nonnegative symmetric solution to (8.14).

To see the relationship between the matrices in (8.14) and the operators in (8.5) more clearly—and the difference between the current approximation scheme and that used in § 4.2 for the control problem—suppose that we take $\hat{Q}_n = P_{En} \hat{Q}|_{En}$. Let $\tilde{\tilde{Q}}^n$ be defined as in (4.36) and (4.37) with Q_0, Q_1, and Q_2 replaced by $\hat{Q}_0$, $\hat{Q}_1$, and $\hat{Q}_2$. Then

$$\tilde{\hat{Q}}^n = W^{-n} \tilde{\tilde{Q}}^n W^{-n}. \tag{8.15}$$

For example, if Q in the control problem and $\hat{Q}$ in the estimator problem are both equal to the identity, then the $\tilde{Q}^n$ in (4.35)-(4.42) is W^n and $\tilde{\hat{Q}}^n = W^{-n}$. This may seem suspicious, but § 8.2 should demonstrate that we are solving the appropriate estimator problem here.

The only thing missing now for numerical implementation of the nth estimator is to give C^n, the matrix representation of C_n, explicitly. We write

$$C^n = [C_1^n \quad C_2^n] \tag{8.16}$$

where the $p \times n$ matrices C_1^n and C_2^n are, respectively, the matrix representations of the approximations to the operators C_1 and C_2 in (7.18). We can cover virtually all applications by assuming $C_n = C|_{En}$, in which case the ith column of C_1^n is the p-vector equal to $C_1 e_i$, and the ith column of C_2^n is the p-vector equal to $C_2 e_i$.

8.2. Stochastic interpretation of the approximating estimators. Using only the deterministic setting above, we will proceed subsequently to analyze the finite-dimensional estimators and the compensators based on them. Nonetheless, we should consider momentarily the sequence of finite-dimensional stochastic estimation problems whose solutions are given by the equations of the preceding section.

First, recall how the covariance operator of a Hilbert space-valued random variable is defined. The covariance operator of an E-valued random variable ω is the operator Q for which

$$\text{expected value } \{\langle z, \omega \rangle_E \langle \hat{z}, \omega \rangle_E\} = \langle Qz, \hat{z} \rangle_E, \quad z, \hat{z} \in E. \tag{8.17}$$

(See [B1], [C4].)

With $\hat{F}_n$ given by (8.4) and (8.5), (8.3) is the Kalman-Bucy filter for the system

$$\dot{z}_n = A_n z_n + B_n u + \omega_n, \tag{8.18}$$

$$y = C_0 u + C_n z_n + \omega_0 \tag{8.19}$$

where $\omega_n(t)$ is an E_n-valued white noise process with covariance operator $\hat{Q}_n$ and $\omega_0(t)$ is an R^p-valued white noise process with covariance operator (matrix) $\hat{R}$. Next, careful inspection will show that the filter defined by (8.6), (8.13), and (8.14) is the matrix representation of the filter defined by (8.3), (8.4), and (8.5).

With z_n and η related as in (4.1) and (4.4), (8.18) and (8.19) are equivalent to the system

$$\dot{\eta} = A^n \eta + B^n u + \nu, \tag{8.20}$$

$$y = C_0 u + C^n \eta + \omega_0 \tag{8.21}$$

where $\nu(t)$ is the R^{2n}-valued noise process related to $\omega_n(t)$ by

$$\omega_n(t) = \sum_{i=1}^{n} (\nu_i(t) e_i,\ \nu_{i+n}(t) e_i). \tag{8.22}$$

Certainly, a Kalman–Bucy filter for (8.20) and (8.21) has the form (8.6) with the filter gain given by (8.13) and (8.14). This particular filter is the matrix representation of the filter defined by (8.3), (8.4), and (8.5) if and only if the matrix $\tilde{\hat{Q}}^n$ defined by (8.12) is the covariance of the process $\nu(t)$. Since $\hat{Q}^n$ is the matrix representation of $\hat{Q}_n$, straightforward calculation using (8.12) and (8.17) shows that the $\tilde{\hat{Q}}^n$ in (8.12) is indeed the correct covariance matrix.

The finite-dimensional observers can be interpreted now as a sequence of filters designed for the sequence of finite-dimensional approximations to the flexible structure, with the nth approximate system disturbed by the noise process $\omega_n(t)$, whose covariance operator is $\hat{Q}_n$. By Hypothesis 8.1, these covariance operators converge to the operator $\hat{Q}$ of § 7. If we have a reliable model of a stationary, zero-mean Gaussian disturbance for the structure, then we can take the covariance operator for this disturbance to be $\hat{Q}$ and think of the infinite-dimensional observer as the optimal estimator. But this interpretation is not necessary for the rest of our analysis.

8.3. The approximating functional estimator gains. The nth estimator gain operator in (8.4) has the same form as the infinite-dimensional estimator gain in (7.15) and (7.20). We have

$$\hat{F}_n y = \sum_{i=1}^{p} (\hat{f}_{in}, \hat{g}_{in}) y_i \tag{8.23}$$

for $y = [y_1 y_2 \cdots y_p]^T \in R^p$, where the functional estimator gains $\hat{f}_{in}$ and $\hat{g}_{in}$ are elements of $V_n = H_n$. The matrix $\hat{F}^n$ in (8.7) and (8.13) is the matrix representation of $\hat{F}_n$. Therefore, if we write

$$\hat{F}^n = \begin{bmatrix} \beta^{f_1} & \beta^{f_2} & \cdots & \beta^{f_m} \\ \beta^{g_1} & \beta^{g_2} & \cdots & \beta^{g_m} \end{bmatrix} \tag{8.24}$$

where the columns $\beta^{f_i}, \beta^{g_i} \in R^n$, then

$$\hat{f}_{in} = \sum_{j=1}^{n} \beta_j^{f_i} e_j, \quad i = 1, \cdots, p, \qquad \hat{g}_{in} = \sum_{j=1}^{n} \beta_j^{g_i} e_j, \quad i = 1, \cdots, p. \tag{8.25}$$

For convergence analysis, it is useful to note from $\hat{f}_{in}$ and $\hat{g}_{in}$ are also given by equations corresponding to (7.22). With the measurement operator C written as in (7.19) and $C_n = C|_E$, we have

$$\hat{f}_{in} = \sum_{j=1}^{p} (\hat{R}^{-1})_{ij} (\hat{\Pi}_{0n} P_{Vn} c_{ij} + \hat{\Pi}_{1n} P_{Hn} c_{2j}), \tag{8.26a}$$

$$\hat{g}_{in} = \sum_{j=1}^{p} (\hat{R}^{-1})_{ij} (\hat{\Pi}_{1n}^{*} P_{Vn} c_{ij} + \hat{\Pi}_{2n} P_{Hn} c_{2j}) \tag{8.26b}$$

where the $\hat{\Pi}_{in}$'s are the blocks of $\hat{\Pi}_n$, as in (7.21).

8.4. Convergence. Now we will indicate the sense in which the finite-dimensional estimators/observers approximate the infinite-dimensional estimator in § 7. The proofs of the convergence results in this section are given in [G6].

THEOREM 8.3. (i) *If $\|\hat{\Pi}_n\|$ is bounded uniformly in n, then the Riccati algebraic equation* (7.16) *has a nonnegative self-adjoint solution $\hat{\Pi}$ and $\hat{\Pi}_n P_{En}$ converges weakly to $\hat{\Pi}$.*

(ii) *If there exist positive constants M and β, independent of n, such that*

$$\|\exp([A_n - \hat{\Pi}_n C_n^* \hat{R}^{-1} C_n]t)\| \leqq M e^{-\beta t}, \qquad t \geqq 0, \tag{8.27}$$

then $\|\hat{\Pi}_n\|$ is bounded uniformly in n, $\hat{\Pi}_n P_{En}$ converges strongly to $\hat{\Pi}$, and $\exp([A_n - \hat{\Pi}_n C_n^* R^{-1} C_n]t) P_{En}$ *converges strongly to $\hat{S}(t)$, the semigroup generated by $A - \hat{\Pi} C^* R^{-1} C$, the convergence uniform in $t \geqq 0$.*

(iii) *If $\hat{Q}_n$ is bounded away from zero uniformly in n, then $\|\hat{\Pi}_n\|$ being bounded uniformly in n guarantees the existence of positive constants M and β for which* (8.27) *holds for all n.*

THEOREM 8.4. *If $\hat{Q}$ is E-coercive and $d_0 = 0$, then there is no nonnegative self-adjoint solution of the Riccati operator equation* (7.16), *and*

$$\|\hat{\Pi}_n\| \to \infty \quad \text{as } n \to \infty. \tag{8.28}$$

Our purpose for bothering to state this dual result to Theorem 5.4 is to point out the following question. Can Theorem 8.4 be modified to include the case where $\hat{Q}$ has the form (7.23) with $\hat{Q}_0 = 0$, $\hat{Q}_1 = 0$, and $\hat{Q}_2$ coercive on H?

Next, we have the dual result to Theorem 5.6.

THEOREM 8.5. *Suppose that A_0 has an invariant subspace V_0 that is also invariant under the damping map D_V, that $E_0 = V_0 \times V_0$ is an observable subspace, and that the restrictions of A_0 and $d_0(\cdot, \cdot)$ to $V_0^\perp$ are both H-coercive. Also, suppose that V_0 has finite dimension n_0 and that, for each $n \geqslant n_0$ in the approximation scheme, the first n_0 e_i's span V_0 and the rest are orthogonal to V_0 in both V and H.*

(i) *Then* (7.16) *has nonnegative solution $\hat{\Pi}$, and $\|\hat{\Pi}_n\|$ is bounded uniformly in n, so that $\hat{\Pi}_n P_{En}$ converges to $\hat{\Pi}$ weakly.*

(ii) *If E_0 and $E_0^\perp$ (the E-orthogonal complement of E_0) are invariant under $\hat{Q}$, and if the E_0-part of the system $(A, \hat{Q})$ is controllable, then the hypothesis of Theorem* 8.3(ii) *holds.*

THEOREM 8.6. *If $\hat{\Pi}_n P_{En}$ converges strongly to $\hat{\Pi}$, then*

$$\|\hat{f}_{in} - \hat{f}_i\|_V \to 0 \quad \text{and} \quad \|\hat{g}_{in} - \hat{g}_i\|_H \to 0, \quad \text{as } n \to \infty \tag{8.29}$$

where $\hat{f}_i$ and $\hat{g}_i$ are the functional estimator gains in (7.20) *and $\hat{f}_{in}$ and $\hat{g}_{in}$ are the approximating functional gains in* (8.25).

9. The finite-dimensional compensators and realizable closed-loop systems.

9.1. Closing the loop. The nth compensator consists of the nth approximation to the optimal control law in § 4, applied to the output of the nth estimator/observer in § 8, i.e., the feedback control

$$u_n = -F_n \hat{z}_n \tag{9.1}$$

where

$$F_n = R^{-1} B_n^* \Pi_n \tag{9.2}$$

(recall (4.25)) and $\hat{z}_n(t)$ is the solution to (8.3). Equivalently, this compensator can be written as

$$u_n = -F^n \hat{\eta} \tag{9.3}$$

where

$$F^n = R^{-1} B^{nT} \tilde{\Pi}^n \tag{9.4}$$

and the $2n$-vector $\hat{\eta}(t)$ is the solution to (8.6). On-line computations will be based on the latter representation, and the block diagram in Fig. 9.1 shows the realizable closed-loop system that results from the nth compensator. We will refer to this system as the nth *closed-loop system.*

This closed-loop system is equivalent to

$$\begin{bmatrix} z(t) \\ \hat{z}_n(t) \end{bmatrix} = S_{\infty n}(t) \begin{bmatrix} z(0) \\ \hat{z}_n(0) \end{bmatrix} \tag{9.5}$$

where $S_{\infty,n}(t)$ is the semigroup on $E \times E_n$ generated by

$$A_{\infty n} = \begin{bmatrix} A & -BF_n \\ \hat{F}_n C & [A_n - B_n F_n - \hat{F}_n C_n] \end{bmatrix}, \qquad D(A_{\infty n}) = D(A) \times E_n. \tag{9.6}$$

Note that $A_{\infty n}$ has compact resolvent if and only if A does.

9.2. Convergence of the closed-loop systems. Now we will consider the sense in which the nth closed-loop system approximates the optimal closed-loop system in § 7 (Definition 7.3). Recall from §§ 4.1 and 8.1 how the approximating open-loop semigroups $T_n(\cdot)$ and their adjoints converge strongly and how the input operators B_n, the measurement operators C_n, and their respective adjoints converge in norm. See [G6] for the proofs of the results in this section.

Hypothesis 9.1. As $n \to \infty$,

$$\|F_n P_{En} - F\| \to 0, \tag{9.7}$$

$$\|\hat{F}_n - \hat{F}\| \to 0. \tag{9.8}$$

Remark 9.2. Of course, we are interested primarily in the case where the gains F and $\hat{F}$ are the optimal LQG gains in (7.15) and (7.17) and F_n and $\hat{F}_n$ are the corresponding approximations in §§ 4 and 8 (i.e., (9.2) and (8.4)). However, for the analysis of this section, we need only Hypothesis 9.1 for some $F \in L(E, R^m)$, $\hat{F} \in L(R^p, E)$ and approximating sequences F_n and $\hat{F}_n$. Any such gain operators will yield closed-loop semigroup generators $A_{\infty\infty}$ in (7.9) and $A_{\infty n}$ in (9.6).

We denote the projection of $E \times E$ onto $E \times E_n$ by P_{EEn}. From the strong convergence of the open-loop semigroups and the uniform norm convergence of the control and estimator gains, we have Theorem 9.3.

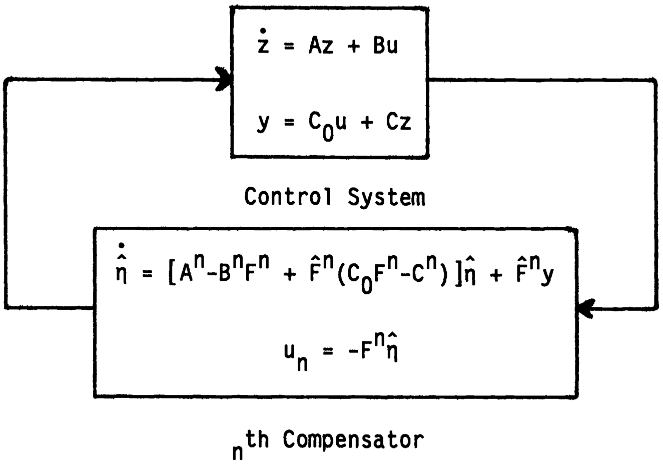


FIG. 9.1. *nth closed-loop system.*

THEOREM 9.3. *For $t \geqq 0$, $S_{\infty n}(t) P_{EEn}$ converges strongly to $S_{\infty\infty}(t)$, and the convergence is uniform in t for t in bounded intervals.*

We should expect at least Theorem 9.3, but we need more. We should require, for example, that if $S_{\infty\infty}(t)$ is uniformly exponentially stable, then $S_{\infty n}(t)$ must be also for n sufficiently large. Although numerical results for numerous examples with various kinds of damping and approximations suggest that this is usually true, we have been unable to prove it in general. We do have the following important case.

THEOREM 9.4. (i) *Suppose that the basis vectors of the approximation scheme are the natural modes of undamped free vibration and that the structural damping does not couple the modes. Then $\| S_{\infty n}(t) - S_{\infty\infty}(t)|_{ExE_n} \|$ converges to zero uniformly in bounded t-intervals.*

(ii) *If, additionally, $S_{\infty\infty}(t)$ is uniformly exponentially stable, then $S_{\infty n}(t)$ is uniformly exponentially stable for n sufficiently large.*

9.3. Convergence of the compensator transfer functions. The transfer function of the nth compensator (shown in the bottom block of Fig. 9.1) is

$$\Phi_n(s) = -F_n(sI - [A_n - B_n F_n + \hat{F}_n(C_0 F_n - C_n)])^{-1} \hat{F}_n, \tag{9.9}$$

which is an $m \times p$ matrix function of the complex variable s for each n, as is the similar transfer function $\Phi(s)$ in (7.14) for the infinite-dimensional compensator. We continue to assume Hypothesis 9.1. The proofs of the following results are given in [G6]. We will denote the resolvent set of $[A - BF + \hat{F}(C_0 F - C)]$ by $\rho([A - BF + \hat{F}(C_0 F - C)])$.

THEOREM 9.5. *There exists a real number a_1 such that, if $\operatorname{Re}(s) > a_1$, then $s \in \rho([A_n - B_n F_n + \hat{F}_n(C_0 F_n - C_n)])$ for all n, and $\Phi_n(s)$ converges to $\Phi(s)$, uniformly in compact subsets of such s.*

This result leaves much to be desired. For example, it does not guarantee that any subset of the imaginary axis will lie in $\rho([A_n - B_n F_n + \hat{F}_n(C_0 F_n - C_n)])$ for sufficiently large n, even if all of the imaginary axis lies in $\rho([A - BF + \hat{F}(C_0 F - C)])$. As with the convergence of the closed-loop systems, we can get more for certain important cases.

Remark 9.6. If the open-loop semigroup $T(\cdot)$ (whose generator is A) is an analytic semigroup, then there exist real numbers a, θ and M, with θ and M positive, such that $\rho([A - BF + \hat{F}(C_0 F - C)])$ contains the sector $\{s\colon |\arg(s - a)| < \pi/2 + \theta\}$, and for each s in this sector,

$$\|(sI - [A - BF + \hat{F}(C_0 F - C)])^{-1}\| \leqq M/|s - a|. \tag{9.10}$$

THEOREM 9.7. (i) *If the basis vectors of the approximation scheme are the natural modes of undamped free vibration and the structural damping does not couple the modes, then each s in $\rho([A - BF + \hat{F}(C_0 F - C)])$ is in $\rho([A_n - B_n F_n + \hat{F}_n(C_0 F_n - C_n)])$ for n sufficiently large and $\Phi_n(s)$ converges to $\Phi(s)$ as $n \to \infty$, uniformly in compact subsets $\rho([A - BF + \hat{F}(C_0 F - C)])$.*

(ii) *If, additionally, $T(\cdot)$ is an analytic semigroup, then $\Phi_n(s)$ converges to $\Phi(s)$ uniformly in the sector described in Remark* 9.6.

THEOREM 9.8. *If A has compact resolvent, then $\Phi_n(s)$ converges to $\Phi(s)$ for each $s \in \rho([A - BF + F(C_0 \hat{F} - C)])$, uniformly in compact subsets.*

10. Closing the loop in the example. As in Definition 7.3, the optimal closed-loop system is formed with the optimal infinite-dimensional compensator, which consists of the optimal control law for the distributed model of the structure applied to the output of an optimal infinite-dimensional state estimator. This optimal control law is the limit of the approximating finite-dimensional control laws in § 6. In this section, we first approximate the infinite-dimensional estimator, as in § 8, and then apply the approximating control laws in § 6 to the approximating finite-dimensional estimators to produce a sequence of finite-dimensional compensators that approximate the optimal compensator.

10.1. The estimator problem. We assume that the only measurement is the rigid-body angle θ and that this measurement has zero-mean Gaussian white noise with variance $\hat{R} = 10^{-4}$. We model the process noise as a zero-mean Gaussian white disturbance that has a component distributed uniformly over the beam, as well as two concentrated components that exert a force on the tip mass and a moment on the hub. For this disturbance, the covariance operator $\hat{Q}$ has the form (7.23) with $\hat{Q}_0 = 0$, $\hat{Q}_1 = 0$, and $\hat{Q}_2 = I$.

We construct the approximating estimators as in § 8.1. The gain for the nth estimator is given by (8.13) with the solution to the Riccati matrix equation (8.14). For the rigid-body measurement, the matrix C^n is

$$C^n = [1 \quad 0 \quad 0 \quad 0 \quad \cdots]. \tag{10.1}$$

According to (8.15), the matrix $\tilde{\hat{Q}}^n$ is

$$\tilde{\hat{Q}}^n = \begin{bmatrix} 0 & 0 \\ 0 & M^{-n} \end{bmatrix}, \tag{10.2}$$

since W^n is the matrix in (4.33). (As always, M^{-n} is the inverse of the mass matrix.) Recall from § 6.3 that $n = 2n_e + 1$ where n_e is the number of elements.

Our only use for the functional estimator gains is to measure the convergence of the finite-dimensional estimators to the optimal infinite-dimensional estimator. To see the convergence of the approximating estimator gains, we compute the approximating functional estimator gains as in § 8.3. As do the functional control gains, the functional estimator gains have the form

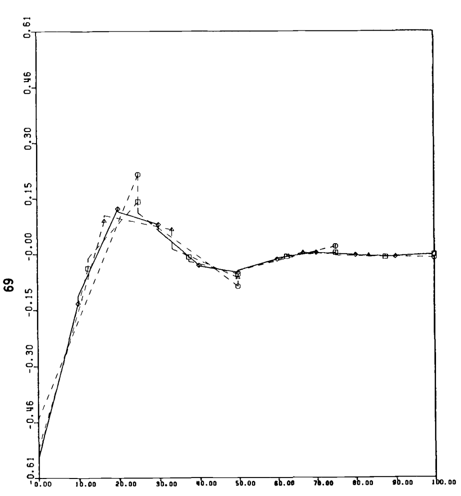
$$\hat{f} = (\alpha_f, \phi_f, \beta_f), \qquad \hat{g} = (\alpha_g, \phi_g, \beta_g), \tag{10.3}$$

and the corresponding approximations have the form

$$\hat{f}_n = (\alpha_{fn}, \phi_{gn}, \beta_{gn}), \qquad \hat{g}_n = (\alpha_{gn}, \phi_{gn}, \beta_{gn}). \tag{10.4}$$

Remark 10.1. We cannot guarantee as much about convergence for the approximating estimators as we could for the approximating control problems in § 6. Since the damping in this example does not couple the natural modes, and the rigid-body mode is observable, we would have part (i) of Theorem 8.5 if we were using the natural mode shapes as basis vectors. Therefore, we know at least that a solution to the infinite-dimensional Riccati equation (7.16) exists and that the infinite-dimensional estimator that we want to approximate exists. The numerical results indicate that the solutions to the finite-dimensional Riccati equations are bounded in n and that the functional estimator gains converge in norm. The rigid-body mode prevents our guaranteeing a priori all the convergence that we want. If a torsional spring and damper were attached to the hub in the current example, we would have coercive stiffness and damping, and Theorem 8.5(ii) would guarantee that the solutions to the finite-dimensional Riccati equations converge strongly and that the functional estimator gains converge in norm for the basis vectors used here.

For damping coefficient $c_0 = 10^{-4}$, Figs. 10.1(a) and 10.1(b) show ϕ''_{fn} and ϕ_{gn}, and Table 10.1 lists the scalars α_{fn}, α_{gn}, and β_{gn}. Since $\phi_{fn}(0) = \phi'_{fn}(0) = 0$, the convergence of ϕ''_{fn} implies the convergence of $\beta_{fn} = \phi_{fn}(l)$; as in the control problem, β_{fn} is



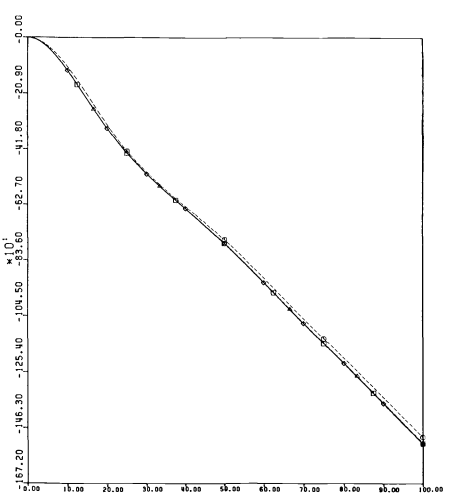


FIG. 10.1(a). *Functional estimator gain component* ϕ''_{fn}. *Damping* $c_0 = 10^{-4}$; *estimator* $R = 10^{-4}$; *number of elements* $n_e = 4, 6, 8, 10$.

FIG. 10.1(b). *Functional estimator gain component* ϕ_{gn}. *Damping* $C_0 = 10^{-4}$; *estimator* $R = 10^{-4}$; *number of elements* $n_e = 4, 6, 8, 10$.

TABLE 10.1
Scalar components of functional estimator gains. Damping coefficient $c_0 = 10^{-4}$; *estimator* $R = 10^{-4}$.

n_e	α_{fn}	α_{gn}	β_{gn}
4	5.3195	14.149	−1495.7
6	5.3567	14.347	−1517.5
8	5.3611	14.371	−1520.1
10	5.3623	14.377	−1520.8

not an independent piece of information about the estimator gains while, as far as our results go, β_{gn} is. We maintain analogy with the control problem and list only β_{gn} in Table 10.1.

10.2. Approximation of the optimal compensator. Finally, for the damping $c_0 = 10^{-4}$, $R = .05$ in the control problem and $\hat{R} = 10^{-4}$ in the estimator problem, we construct the finite-dimensional compensator in Fig. 9.1; i.e., for each $n = 2n_e + 1$, we apply the n_eth control law represented by the functional gains in Fig. 6.2 and Table 6.2 to the output of the n_eth estimator represented by the functional gains in Fig. 10.1 and Table 10.1. As the number of elements increases, the transfer function in (9.9) of the finite-dimensional compensator converges to the transfer function in (7.14) of the optimal infinite-dimensional compensator, as described in § 9.3. Theorem 9.5 and Remark 9.6 apply. Figure 10.2 shows the frequency response (bode plots) of the finite-dimensional compensators for 4, 6, 8, and 10 elements. The phase plot is for 10 elements only. These plots indicate that the finite-dimensional compensator for eight or more elements is virtually identical to the optimal infinite-dimensional compensator, as predicted by the functional gain convergence in Figs. 6.2 and 10.1

10.3. Comments on the structure and dimension of the implementable compensators. Although this paper does not address the problem of obtaining the lowest-order compensator that closely approximates the infinite-dimensional compensator,

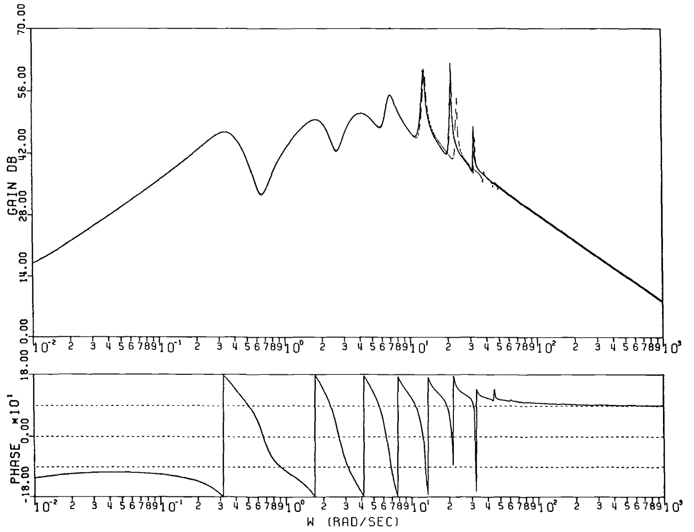


FIG. 10.2. *Frequency response* (*bode plot*) *of compensators. Damping* $c_0 = 10^{-4}$; *control* $R = .05$, *estimator* $R = 10^{-4}$; *number of elements* $n_e = 4$, 6, 8, 10.

we should note that the compensators based on eight and ten elements here are unnecessarily large because the finite-element scheme that we chose is not nearly the most efficient in terms of the dimension required for convergence. (The dimension of the first-order differential equation in the compensator is $2(2n_e + 1)$.) We used cubic Hermite splines here to demonstrate that the finite-element scheme most often used to approximate beams in other engineering applications can be used in approximating the optimal compensator. In [G5], we compare the present scheme with one using cubic B-splines and one using the natural mode shapes as basis vectors. The natural mode shapes yield the fastest converging compensators, but the B-splines are almost as good. The only advantages of the Hermite splines result from the fact that the coding to build the basic matrices (mass, stiffness, etc.) is simpler than for B-splines and the fact that, before the Riccati equations based on, say, 10 natural modes are solved, a much larger finite-element approximation of the structure must be used to get the 10 modes accurately.

To understand the redundancy in the large finite-dimensional compensators here, it helps to consider the structure of the optimal compensator. It is based on an infinite-dimensional state estimator that has a representation of each of the structure's modes. In the present example, the optimal compensator estimates and controls the first six modes significantly, the next three modes slightly, and virtually ignores the rest. This observation is based on the projections of the functional gains onto the natural modes and on comparison of the open-loop and closed-loop eigenvalues. (See [G5] for more detail, including the spectrum of the closed-loop system—which is stable—obtained with the ten-element compensator here.) The infinite-dimensional compensator then has an infinite number of modes that contribute nothing to the input-output map of the compensator. These inactive modes are just copies of all the open-loop modes past the first nine. They can be truncated from the compensator without affecting the closed-loop system response significantly. The number of active modes in the compensator—i.e., the modes that contribute to the input-output map—depends on the structural damping and the Q's and R's in the LQG problem statement.

(See the discussion in § 6.4 about the effect of damping and control weighting on performance.)

The compensator computed here based on 10 elements has 21 modes (although we did not do the computations in modal coordinates). Nine of these compensator modes are virtually identical to the nine active modes in the infinite-dimensional compensator, and the 12 inactive modes are approximations to the tenth through twenty-first open-loop modes of the structure. The inactive modes result from the large number of elements needed to approximate the active compensator modes accurately. Now that we essentially have the optimal compensator in the ten-element compensator, we could truncate the 12 inactive modes and implement a compensator with nine modes. And we probably could reduce the compensator even further using an order reduction method such as balanced realizations.

11. Conclusions. For the deterministic linear-quadratic optimal regulator problem for a flexible structure with bounded input operator (the B_0 in (2.1)), the approximation theory in §§ 4 and 5 is reasonably complete. The most important extensions should be to the corresponding (very difficult) problem with unbounded input operator, for which there exists little approximation theory. Because of the different kinds of boundary input operators, stiffness operators and structural damping, all of which must be considered in detail when B_0 is unbounded, it seems unlikely that the approximation theory for the unbounded-input case can be made as complete as the theory here.

The convergence results in § 8 for the estimation problem are less complete than those for the control problem because rigid-body modes present more technical difficulties for the proofs in the estimator case. However, our analysis and numerical experience suggest that the difficulties only make the proofs harder and that the convergence in the estimation problem is identical to the convergence in the control problem, and that controllable and observable rigid-body modes make no qualitative difference in either problem.

Where we would most like to have substantial improvement over the results of this paper is in § 9.2, which considers how the approximating closed-loop systems obtained by controlling the distributed model of the structure with the finite-dimensional compensators converge to the optimal closed-loop system, obtained with the infinite-dimensional compensator. Theorem 9.4 gives us what we want for problems where the damping does not couple the natural modes of free vibration, and where the natural mode shapes are the basis vectors for the approximation scheme. In particular, this theorem says that, if the optimal closed-loop system is uniformly exponentially stable, then so are the approximating closed-loop systems for sufficiently large order of approximation. We have verified numerically the stability of the approximating closed-loop systems for the example in §§ 6 and 10 where the basis vectors are not the modes. This example and others have made us suspect that Theorem 9.4 is true when the basis vectors satisfy Hypothesis 4.1 only and when the damping couples the modes. The methods in [I1] should be useful in completing the analysis.

Another possible approach to analyzing the convergence of the approximating closed-loop systems to the optimal closed-loop system is to use the input–output description in frequency domain. Results such as those in § 9.3 are useful for this, although for the closed-loop stability we want, we probably need the transfer functions of the finite-dimensional compensators to converge more uniformly on the compensator resolvent set than we have proved here. In our example, Fig. 10.2 indicates that these transfer functions converge uniformly on the imaginary axis, but we have no theorem that guarantees this.

Appendix. Errata for [G1]. In the first paragraph of the proof of Theorem 2.1 on page 689 of [G1], the first sentence should be:

> If a dissipative operator is invertible, its inverse is dissipative.

At the beginning of the fifth line of the same paragraph, the expression $(\alpha x + y)$ should be deleted the first time it occurs. The next-to-last sentence of the paragraph should be:

> Hence, if a densely defined maximal dissipative operator has dense range, its inverse is maximal dissipative.

The theorem is correct as stated.

In the current paper, we use Theorem 2.1 of [G1] to conclude that the operator $\tilde{A}$ defined in § 2 is maximal dissipative (see (2.10)-(2.12)).

REFERENCES

[B1] A. V. BALAKRISHNAN, *Applied Functional Analysis*, 2nd ed., Springer-Verlag, New York, 1981.

[B2] ———, *Strong stability and the steady-state Riccati equation*, Appl. Math. Optim., 7 (1981), pp. 335-345.

[B3] M. J. BALAS, *Feedback control of flexible systems*, IEEE Trans. Automat. Control, (1978), pp. 673-679.

[B4] ———, *Modal control of certain flexible systems*, SIAM J. Control Optim., 16 (1978), pp. 450-462.

[B5] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: numerical methods based on averaging approximations*, SIAM J. Control Optim., 16 (1978), pp. 169-208.

[B6] H. T. BANKS AND K. KUNISCH, *The linear regulator problem for parabolic systems*, SIAM J. Control Optim., 22 (1984), pp. 684-698.

[B7] D. S. BERNSTEIN AND D. C. HYLAND, *The optimal projection approach to designing finite-dimensional controllers for distributed parameter systems*, in Proc. 23rd IEEE Conference on Decision and Control, Las Vegas, NV, December 1984.

[B8] ———, *The optimal projection equations for finite-dimensional fixed-order dynamic compensation of infinite-dimensional systems*, SIAM J. Control Optim., to appear.

[C1] G. CHEN AND D. L. RUSSELL, *A mathematical model for linear elastic systems with structural damping*, Quart. Appl. Math., (1982), pp. 433-454.

[C2] R. W. CLOUGH AND J. PENZIEN, *Dynamics of Structures*, McGraw-Hill, New York, 1975.

[C3] R. F. CURTAIN, *On stabilizability of linear spectral systems via state boundary feedback*, SIAM J. Control Optim., 23 (1985), pp. 144-152.

[C4] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Springer-Verlag, New York, 1978.

[C5] R. F. CURTAIN AND D. SALAMON, *Finite dimensional compensators for infinite dimensional systems with unbounded control action*, to appear.

[C6] S. CHANG AND I. LASIECKA, *Riccati equations for nonsymmetric and nondissipative hyperbolic systems with L_2-boundary controls*, J. Math. Anal. Appl., 116 (1986), pp. 378-414.

[D1] N. DUNFORD AND J. T. SCHWARTZ, *Linear Operators, Part* II, Wiley Interscience, New York, 1963.

[D2] G. DA PRATO, I. LASIECKA, AND R. TRIGGIANI, *A direct study of the Riccati equation arising in hyperbolic boundary control problems*, J. Differential Equations, 64 (1986), pp. 26-47.

[G1] J. S. GIBSON, *An analysis of optimal model regulation: convergence and stability*, SIAM J. Control Optim., 19 (1981), pp. 686-707.

[G2] ———, *A note on stabilization of infinite dimensional linear oscillators by compact linear feedback*, SIAM J. Control Optim., 18 (1980), pp. 311-316.

[G3] ———, *Linear-quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95-139.

[G4] ———, *The Riccati integral equations for optimal control problems on Hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537-565.

[G5] J. S. GIBSON AND A. ADAMIAN, *A comparison of approximation schemes for optimal control of flexible structures*, to appear.

[G6] ———, *Approximation theory for* LQG *optimal control of flexible structures*, ICASE Report No. 88-48, Institute for Computer Applications in Science and Engineering, NASA Langley Research Center, Hampton, VA, 1988.

[G7] J. S. GIBSON AND D. L. MINGORI, *Approximation and compensator order determination in optimal control of flexible structures*, in Proc. IFAC Workshop on Model Error Concepts and Compensation, Boston, MA, June 1985.

[G8] J. S. GIBSON, D. L. MINGORI, A. ADAMIAN, AND F. JABBARI, *Approximation of optimal infinite dimensional compensators for flexible structures*, JPL Workshop on Identification and Control of Space Structures, Jet Propulsion Laboratory, San Diego, CA, June 1984.

[G9] J. S. GIBSON, D. L. MINGORI, A. ADAMIAN, AND P. A. BILELLOCH, *Integrated Control/Structure Research for Large Space Structures*, HR Textron Report #956541-Final, to Jet Propulsion Laboratory, September 1984.

[G10] J. S. GIBSON AND M. NAVID, *Approximate solution of Riccati algebraic equations in optimal control and estimation of hyperbolic systems*, International Symposium on Mathematical Theory of Networks and Systems, Santa Monica, CA, August 1981.

[I1] K. ITO, *Finite dimensional compensators for infinite dimensional systems via Galerkin-type approximations*, SIAM J. Control Optim., 28 (1990), pp. 1251-1269.

[I2] K. ITO AND R. POWERS, *Chandrasekhar equations for infinite dimensional systems* II, *Unbounded input and output case*, J. Differential Equations, 75 (1988), pp. 371-402.

[I3] K. ITO AND H. T. TRAN, *Linear optimal control problem for linear systems with unbounded input and output operators*, numerical approximations, in International Series of Numerical Mathematics, Vol. 91, Birkhauser Verlag, 1989, pp. 171-195.

[K1] T. KATO, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, New York, 1984.

[L1] I. LASIECKA AND R. TRIGGIANI, *Dirichlet boundary control problem for parabolic equations with quadratic cost: analyticity and Riccati's feedback synthesis*, SIAM J. Control Optim., 21 (1983), pp. 41-67.

[L2] ———, *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. 884-925.

[L3] ———, *The regulator problem for parabolic equations with Dirichlet boundary control*, Parts I and II, Appl. Math. Optim., 16 (1987), pp. 147-168, 187-216.

[L4] I. LASIECKA, *Approximations of Riccati equation for abstract boundary control problems—applications to hyperbolic systems*, Numer. Funct. Anal. Optim., 8 (1985-86), pp. 207-243.

[M1] D. L. MINGORI, J. S. GIBSON, P BLELLOCH, AND A. ADAMIAN, *Control of a flexible space antenna: a finite dimensional perspective based on distributed parameter theory*, JPL Workshop on Identification and Control of Space Structures, Jet Propulsion Laboratory, San Diego, CA, June 1984.

[M2] B. C. MOORE, *Principal component analysis in linear systems: controllability, observability, and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17-32.

[S1] M. H. SCHULTZ, *Spline Analysis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[S2] J. M. SCHUMACHER, *A direct approach to compensator design for distributed parameter systems*, SIAM J. Control Optim., 21 (1983), pp. 823-836.

[S3] R. E. SHOWALTER, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, 1977.

[S4] G. STRANG AND G. J. FIX, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 1, pp. 38-57, January 1991

SUFFICIENT CONDITIONS FOR DYNAMIC STATE FEEDBACK LINEARIZATION*

B. CHARLET†, J. LÉVINE†, AND R. MARINO‡

Abstract. Sufficient conditions are given for the existence of a dynamic state feedback compensator for a multi-input nonlinear system such that the closed loop system is transformable into a linear controllable one by an extended state space change of coordinates. An example shows that the conditions are not necessary. Necessary conditions are also given which are shown to be sufficient when the number of states minus the number of controls is equal to one. Several examples illustrate how the sufficient conditions obtained lead to the design of the dynamic compensator.

Key words. nonlinear systems, feedback linearization, dynamic compensator

AMS(MOS) subject classifications. 93B10, 93C10, 53A55

1. Introduction. We address the problem of transforming a nonlinear control system

$$\dot{z} = f(z) + \sum_{i=1}^{m} g_i(z)u_i(t) = f(z) + G(z)u \qquad z \in \mathbb{R}^n, \quad u \in \mathbb{R}^m \tag{1}$$

with $f(0) = 0$ and rank $G(0) = m$ into a linear controllable system

$$\dot{x} = Ax + Bv \qquad x \in \mathbb{R}^{n'}, \quad v \in \mathbb{R}^{m'} \tag{2}$$

with $n' \geqq n$, $m' \geqq m$.

Since 1973 [16] this problem has been studied using increasingly more general transformations. State space diffeomorphisms

$$x = \varphi(z), \quad \varphi(0) = 0, \quad x \in \mathbb{R}^n, \quad z \in \mathbb{R}^n \tag{3}$$

were the first transformations to be studied in [16]. State feedback transformations

$$u = \alpha(z) + \beta v \qquad v \in \mathbb{R}^m \tag{4}$$

with $\alpha(0) = 0$ and β a nonsingular $m \times m$ constant matrix were then introduced by Brockett [3] and later generalized in [15] and [11] by

$$u = \alpha(z) + \beta(z)v \qquad v \in \mathbb{R}^m \tag{5}$$

where the nonsingular matrix β was allowed to depend on the state as well. The study of the above transformations led to complete characterizations of those systems (1) transformable into (2) by (3) ([16], [24]), which are called state linearizable, and of those systems transformable into (2) by (3) and (5) ([15], [11]), which are called static feedback linearizable. Adaptive feedback linearization and its robustness versus unmodeled dynamics are studied in [25]. For those systems that are not static feedback

* Received by the editors, April 4, 1988; accepted for publication (in revised form) December 13, 1989. A preliminary version of this paper [6] was presented at the International Federation of Automatic Control Conference "Nonlinear Control Systems Design," in Capri, June 14-16, 1989.

† Section Automatique-C.A.I., École Nationale Supérieure des Mines de Paris, 35 rue Saint Honoré, 77305 Fontainebleau Cédex, France. B. Charlet is on leave from Délégation Générale de l'Armement, Minstère de la Défense Nationale.

‡ Dipartimento di Ingegneria Elettronica, Seconda Università di Roma "Tor Vergata," Via O. Raimondo, 00173 Rome, Italy. The research of this author was partially supported by Ministero Pubblica Istruzione (Fondi 40%).

linearizable, the problem of partial feedback linearization was posed in [17], namely, the transformation of (1) by (3) and (5) into a partially linear and controllable system

$$\begin{aligned} \dot{x}^{(1)} &= Ax^{(1)} + Bv && x^{(1)} \in \mathbb{R}^{p} \\ \dot{x}^{(2)} &= \gamma(x^{(1)}, x^{(2)}) + \delta(x^{(1)}, x^{(2)})v && x^{(2)} \in \mathbb{R}^{n-p} \end{aligned} \tag{6}$$

where (A, B) is a controllable pair. This problem was solved in [17], [18] where the dimension and the construction of the largest feedback linearizable subsystem are given for any system (1).

Partial feedback linearization is related to input-output decoupling. Given m outputs

$$y_j = h_j(z) \qquad 1 \leqq j \leqq m \tag{7}$$

the input-output decoupling problem is to determine a transformation (3), (5) that takes the system (1), (7) into

$$\begin{aligned} \dot{x}^{(1)} &= Ax^{(1)} + Bv && x^{(1)} \in \mathbb{R}^{p} \\ \dot{x}^{(2)} &= \gamma(x^{(1)}, x^{(2)}) + \delta(x^{(1)}, x^{(2)})v && x^{(2)} \in \mathbb{R}^{n-p} \\ y &= Cx^{(1)} && \end{aligned} \tag{8}$$

with (A, B, C) in prime canonical form [19], namely, $A = \text{block diag}\,[A_1, \cdots, A_p]$, $B = \text{block diag}\,[B_1, \cdots, B_p]$, $C = \text{block diag}\,[C_1, \cdots, C_p]$, with, for every $i = 1, \cdots, p$,

$$A_i = \begin{pmatrix} 0 & 1 & & \\ & \ddots & \ddots & 0 \\ 0 & & \ddots & \ddots \\ & & \ddots & 1 \\ & & & 0 \end{pmatrix}, \quad B_i = \begin{pmatrix} 0 \\ \vdots \\ \vdots \\ 0 \\ 1 \end{pmatrix}, \quad C_i = (0 \quad \cdots \quad \cdots \quad 0 \quad 1).$$

Necessary and sufficient conditions are available for this problem (see [13]). For linear systems it is known that those conditions can be weakened if one allows for a dynamic compensator [20]. This motivated the introduction of a nonlinear dynamic state feedback transformation [23]

$$\begin{aligned} \dot{w} &= a(z, w) + B(z, w)v && w \in \mathbb{R}^{q} \\ u &= \alpha(z, w) + \beta(z, w)v && v \in \mathbb{R}^{m} \end{aligned} \tag{9}$$

with $\alpha(0, 0) = 0$, $a(0, 0) = 0$. The dynamic state feedback (9) is a generalization of the static state feedback (5).

Necessary and sufficient conditions for input-output decoupling via transformations (9) and extended state space diffeomorphism

$$x = \varphi(z, w) \qquad x \in \mathbb{R}^{n+q} \tag{10}$$

were obtained in [7], [8], and [21]. In [12] and [14] sufficient conditions are given to achieve both input-output decoupling and full linearization, that is, $x^{(2)} = 0$ in (6), via transformations (9), (10).

In [4] and [5] the problem of transforming the system (1) into (2) via transformations (9) and (10), which will be called the dynamic feedback linearization problem, is studied: it is shown that single input systems (1) that are dynamically feedback linearizable are also statically feedback linearizable; two very special classes of dynamically feedback linearizable multi-input systems are given in [4]. A different approach is taken by [9]: transformations (9) are considered, but the closed-loop system is no longer required to be linear and controllable in some new coordinates of the extended space (z, w): the only requirement is the existence of a w-dependent state space change of coordinates

$$x=\varphi(z, w), \qquad \varphi(0,0)=0, \qquad z\in\mathbb{R}^n, \quad x\in\mathbb{R}^n \tag{11}$$

in which the system is linear and controllable from v. While this is a less restrictive notion of dynamic feedback linearization, the stability properties of the dynamic compensator (9) remain to be analyzed. This analysis, which could be quite difficult, is not needed with our notion of dynamic feedback linearization.

In this paper we present fairly general sufficient conditions for a system (1) to be dynamic feedback linearizable by a special class of dynamic compensators (prolongations) and extended space diffeomorphism (10). Compensators are restricted to be of the following form ($u^{(\mu)}=(d^\mu u/dt^\mu)$, $w=(u_1,\cdots,u_1^{(\mu_1)},\cdots,u_m,\cdots,u_m^{(\mu_m)})$):

$$\begin{pmatrix} u_1^{(\mu_1)} \\ \vdots \\ u_m^{(\mu_m)} \end{pmatrix} = \alpha(z,w)+\beta(z,w)\begin{pmatrix} v_1 \\ \vdots \\ v_{m'} \end{pmatrix} \tag{12}$$

with $\mu_i\geqq 0$, $1\leqq i\leqq m$, $\mu=\sum_{i=1}^m \mu_i$, $\alpha(0,0)=0$, $\beta(z,w)$ of full rank m ($m'\geqq m$) in V_0, a neighborhood of the origin in $\mathbb{R}^{n+\mu}$, which can be realized as

$$\begin{aligned}
&\dot w_i^j = w_{i+1}^j \qquad 1\leqq i\leqq \mu_j-1, \quad 1\leqq j\leqq m, \quad \mu_j>1\\
&\dot w_{\mu_j}^j=\alpha_j(z,w)+\sum_{l=1}^{m'}\beta_{j,l}(z,w)v_l(t)=v_j'(t) \qquad 1\leqq j\leqq m, \quad \mu_j>0\\
&u_j=w_1^j \qquad 1\leqq j\leqq m, \quad \mu_j>0\\
&u_j=\alpha_j(z,w)+\sum_{l=1}^{m'}\beta_{j,l}(z,w)v_l(t)=v_j'(t) \qquad 1\leqq j\leqq m, \quad \mu_j=0.
\end{aligned} \tag{13}$$

In § 4 sufficient conditions are obtained that require the involutivity of certain distributions defined on the original state space. The distributions are computed on the basis of the given vector fields $f, g_1, \cdots, g_m$ and on a set of integers $\mu_1, \cdots, \mu_m$ that characterize the linearizing dynamic compensator (13). The sufficient conditions are helpful in determining the structure of the dynamic compensator on the basis of the Lie algebraic structure of the system. When $\mu_1=\cdots=\mu_m=0$ and $m'=m$, the conditions coincide with the necessary and sufficient conditions for static feedback linearization. We show by examples that the conditions are not necessary. However, they are general enough to apply for a detailed nonlinear model of rigid body dynamics that is not static feedback linearizable but is shown to be dynamically feedback linearizable.

In § 3 we show that the controllability of the linear approximation of system (1) at the origin is a necessary condition for dynamic feedback linearization; as a corollary of the main result of § 4, it turns out that it is also sufficient when $m=n-1$ in (1).

2. Basic results and definitions. In this section, we consider a general dynamic compensator of the form

$$\begin{aligned}
\dot w &= a(z,w)+B(z,w)v \qquad w\in\mathbb{R}^q\\
u &= \alpha(z,w)+\beta(z,w)v \qquad v\in\mathbb{R}^{m'}, \quad m'\geqq m.
\end{aligned} \tag{14}$$

System (1) controlled by (14), which is called the extended system, becomes

$$\begin{aligned}
\dot z &= f(z)+G(z)\alpha(z,w)+G(z)\beta(z,w)v\\
\dot w &= a(z,w)+B(z,w)v
\end{aligned} \tag{15}$$

and is written as

$$\dot{\tilde{z}} = \tilde{f}(\tilde{z}) + \tilde{G}(\tilde{z})v = \tilde{f}(\tilde{z}) + \sum_{i=1}^{m'} \tilde{g}_i(\tilde{z})v_i \tag{16}$$

where $\tilde{z} = (z, w)$ is the extended state and

$$\tilde{f} = \begin{pmatrix} f(z) + G(z)\alpha(z, w) \\ a(z, w) \end{pmatrix}, \qquad \tilde{G} = \begin{pmatrix} G(z)\beta(z, w) \\ B(z, w) \end{pmatrix}. \tag{17}$$

If $u_j = \alpha_j(\tilde{z}) + \sum_{i=1}^{m'} \beta_{j,i}(\tilde{z})v_i$, $1 \leqq j \leqq m$, are viewed as m outputs for system (16), we can define the corresponding differential output rank (see [1], [10]). Denoting by $u^{(k)}$ the kth time derivative of u considered as an output for system (16), we define the sequence of subspaces $E_0 \subset E_1 \subset \cdots \subset E_{n+q}$ by:

$$E_k = sp\{\mathrm{d}\tilde{z}, \mathrm{d}\dot{u}, \cdots, \mathrm{d}u^{(k)}\}.$$

The differential output rank $d^o(u)$ is defined by:

$$d^o(u) = \dim E_{n+q} - \dim E_{n+q-1}.$$

In the single input case ($m = 1$), it is proved in [1] that the computation of the differential output rank reduces to the computation of the classical rank of the decoupling matrix (see, for example, [13]).

In this case, the characteristic index ν is defined by:

$$\begin{aligned} &\nu = 0 \quad \text{if } \beta_i(\tilde{z}) \neq 0 \qquad \text{for some } i, \quad 1 \leqq i \leqq m' \\ &\nu = \min\{r \mid L_{\tilde{g}_i} L_{\tilde{f}}^{r-1}\alpha(\tilde{z}) \neq 0 \quad \text{for some } i, \quad 1 \leqq i \leqq m'\}.^{(1)} \end{aligned} \tag{18}$$

Let

$$\delta_i(\tilde{z}) = \begin{cases} \beta_i(\tilde{z}) & \text{if } \nu = 0 \\ L_{\tilde{g}_i} L_{\tilde{f}}^{\nu-1} a(\tilde{z}) & \text{if } \nu > 0. \end{cases} \tag{19}$$

When ν is finite, the $1 \times m'$ matrix

$$D(\tilde{z}) = (\delta_1(\tilde{z}), \cdots, \delta_{m'}(\tilde{z})) \tag{20}$$

is called the decoupling matrix of the compensator (14) for system (1) with $m = 1$. In this case, $d^o(u) = \operatorname{rank} D$.

DEFINITION 2.1. A dynamic compensator (14) is said to be regular for system (1) if the corresponding differential output rank $d^o(u)$ is equal to m in V_0, a neighborhood of the origin in $\mathbb{R}^{n+q}$.

Note that the static state feedback (5), with $\beta(z)$ nonsingular, or the dynamic compensator (13) with $\beta(\tilde{z})$ nonsingular, are nonsingular compensators according to the above definition.

DEFINITION 2.2. A system (1) is said to be locally static feedback linearizable if there exists a feedback transformation (5), (3), in a neighborhood of $z = 0$, which transforms system (1) into system (2) with $n' = n$, $m' = m$.

DEFINITION 2.3. A system (1) is said to be locally dynamic feedback linearizable if there exists a regular dynamic compensator (14) and an extended space diffeomorphism (10) defined in a neighborhood of $(z, w) = (0, 0)$ which transform system (1) into system (2) with $n' = n + q$, and such that $\varphi(0, 0) = (0, 0)$.

[1] If f is a smooth vector field and α a smooth function, the *Lie derivative of* α *with respect to* f is defined, in local coordinates, by $L_f\alpha = \sum_{n=1}^{n} f_i(\partial\alpha/\partial x_i)$. We also not $L_f^0\alpha = \alpha$ and $L_f^k\alpha = L_f(L_f^{k-1}\alpha)$ for all $k \geqq 1$.

We now recall from [15] and [11] basic definitions and results. Define the distributions

$$\mathcal{G}_0 = sp\{g_j, 1 \leqq j \leqq m\} \tag{21}$$
$$\mathcal{G}_i = sp\{\mathrm{ad}_f^l g_j, 0 \leqq l \leqq i, 1 \leqq j \leqq m\},^{(2)} \qquad i > 0,$$

which enjoy the recursion properties

$$\mathcal{G}_{i+1} = \mathcal{G}_i + \mathrm{ad}_f^{i+1}\mathcal{G}_0 = \mathcal{G}_i + \mathrm{ad}_f\mathcal{G}_i \tag{22}$$

where $\mathrm{ad}_f\mathcal{G} = sp\{\mathrm{ad}_f\, Y, Y \in \mathcal{G}\}$. By definition $\mathcal{G}_0 \subset \mathcal{G}_1 \subset \cdots \subset \mathcal{G}_i \subset \cdots$.

We recall that a distribution $\mathcal{G}$ is involutive if and only if, given any pair of vector fields g_1 and g_2 in $\mathcal{G}$, their Lie bracket $[g_1, g_2]$ belongs to $\mathcal{G}$.

THEOREM 2.1. *System* (1) *is locally static feedback linearizable if and only if, in* U_0, *a neighborhood of the origin in* $\mathbb{R}^n$:

(i) $\mathcal{G}_i$ *is an involutive distribution of constant rank for every* $i \geqq 0$;

(ii) rank $\mathcal{G}_{n-1} = n$.

The distributions $\mathcal{G}_i$ are invariant under change of coordinates (3), and under the assumption (i), they are invariant under state feedback transformations (5).

Let $m_0 = \mathrm{rank}\, \mathcal{G}_0$, $m_i = \mathrm{rank}\, \mathcal{G}_i - \mathrm{rank}\, \mathcal{G}_{i-1}$, $i > 1$, in U_0. The m indices (since $m_0 = m$),

$$k_j = \mathrm{card}\,\{i \mid m_i \geqq j\}, \qquad 1 \leqq j \leqq m,$$

are uniquely associated with a system (1) satisfying (i) and (ii): they are invariant under a state-space change of coordinates (3) and state feedback transformations (5). They are called controllability indices.

It has been shown in [5] that dynamic feedback linearizability implies static feedback linearizability when $m' = m = 1$. We now generalize this result for arbitrary $m' \geqq m = 1$.

THEOREM 2.2. *The following statements are equivalent*:

(i) *System* (1) *with* $m = 1$ *is locally static feedback linearizable.*

(ii) *System* (1) *with* $m = 1$ *is locally dynamic feedback linearizable.*

Proof. The proof is an easy generalization of the main result in [5]; (i)⇒(ii) is obvious.

(ii)⇒(i). By assumption (ii) there exists a linearizing compensator (9) for system (1) with $m = 1$ and characteristic index ν. We first establish a relationship between the distributions $\mathcal{G}_i$ defined by (21) for system (1) and the distributions $\tilde{\mathcal{G}}_i$ defined for system (16) with $m = 1$ obtained by using the linearizing compensator:

$$\tilde{\mathcal{G}}_0 = sp\{\tilde{g}_k, 1 \leqq k \leqq m'\} \tag{23}$$
$$\tilde{\mathcal{G}}_i = \tilde{\mathcal{G}}_{i-1} + sp\{\mathrm{ad}_{\tilde{f}}^i\, \tilde{g}_k, 1 \leqq k \leqq m'\}$$

CLAIM.

$$\mathrm{ad}_{\tilde{f}}^i\, \tilde{g}_j = \begin{pmatrix} 0 \\ * \end{pmatrix} \qquad \text{if } i < \nu \tag{24}$$
$$\mathrm{ad}_{\tilde{f}}^i\, \tilde{g}_j = \begin{pmatrix} \gamma_j \mathrm{ad}_f^{i-\nu}\, g + X_{j,i} \\ * \end{pmatrix} \qquad \text{if } i \geqq \nu,$$

where $X_{j,i} \in \overline{\mathcal{G}_{i-\nu-1}}$ *and* $\gamma_j = (-1)^\nu \delta_j$, δ_j *being defined by* (19).

[2] We denote by $\mathrm{ad}_f\, g = [f, g]$ the *Lie bracket of the smooth vector fields f and g.* In local coordinates, $[f, g] = \sum_{n=1}^n \sum_{j=1}^n f_j(\partial g_i/\partial x_j) - g_j(\partial f_i/\partial x_j))(\partial/\partial x_i)$. We also note $\mathrm{ad}_f^0\, g = g$ and $\mathrm{ad}_f^k\, g = [f, \mathrm{ad}_f^{k-1}\, g]$ for all $k \geqq 1$.

Proof of the claim. The proof is by induction on i. For $i=0$ we have

$$\tilde{g}_j=\begin{pmatrix}\beta_j g\\ b\end{pmatrix}$$

with $\beta_j=\gamma_j$ and $X_{j,0}=0$ if $\nu=0$ and $\beta_j=0$ if $\nu\geqq 1$.

We assume that the claim holds true for an integer i: we shall prove it for $i+1$. We consider two cases: $i\geqq\nu$; $i<\nu$. If $i\geqq\nu$,

$$\begin{aligned}\mathrm{ad}_f^{i+1}\tilde{g}_j&=\left[\begin{pmatrix}f(z)+\alpha(z,w)g(z)\\ a(z,w)\end{pmatrix},\begin{pmatrix}\gamma_j(z,w)\,\mathrm{ad}_f^{i-\nu}g+X_{j,i}\\ *\end{pmatrix}\right]\\ &=\begin{pmatrix}\gamma_j\,\mathrm{ad}_f^{i+1-\nu}g+X_{j,i+1}\\ *\end{pmatrix},\end{aligned}$$

where

$$X_{j,i+1}=[f,X_{j,i}]+\alpha[g,X_{j,i}]+(L_{\tilde f}\gamma_j)\,\mathrm{ad}_f^{i-\nu}g-(L_{\mathrm{ad}_f^i}\tilde{g}\alpha)g+\alpha\gamma_j\,\mathrm{ad}_g\,\mathrm{ad}_f^{i-\nu}g+\frac{\partial X_{j,i}}{\partial w}a.$$

Since $X_{j,i}\in\overline{\mathscr{G}_{i-\nu-1}}$ and $\gamma_j=(-1)^\nu\delta_j$, $X_{j,i}$ can be written as

$$X_{j,i}(z,w)=\sum_k\sigma_k(z,w)\Gamma_{k,j,i}(z),$$

with $\Gamma_{k,j,i}\in\overline{\mathscr{G}_{i-\nu-1}}$, thus

$$\frac{\partial X_{j,i}}{\partial w}a=\sum_k(L_a\sigma_k)\Gamma_{k,j,i}\in\overline{\mathscr{G}_{i-\nu-1}}$$

and $X_{j,i+1}$ belongs to $\overline{\mathscr{G}_{i-\nu}}$.

If $i<\nu$, by assumption

$$\mathrm{ad}_{\tilde f}^i\tilde{g}_j=\begin{pmatrix}0\\ *\end{pmatrix}$$

so

$$\mathrm{ad}_{\tilde f}^{i+1}\tilde{g}_j=\begin{pmatrix}-(L_{\mathrm{ad}_f^i}\tilde{g}_j\alpha)g\\ *\end{pmatrix}.$$

Since

$$L_{\mathrm{ad}_f^i}\tilde{g}_j\alpha=\begin{cases}0 & \text{if } i<\nu-1\\ (-1)^iL_{\tilde{g}_j}L_{\tilde f}^i\alpha & \text{if } i=\nu-1\end{cases}$$

then

$$\mathrm{ad}_{\tilde f}^{i+1}\tilde{g}_j=\begin{cases}\begin{pmatrix}0\\ *\end{pmatrix} & \text{if } i\leqq\nu-2\\ \begin{pmatrix}\gamma_j g\\ *\end{pmatrix} & \text{if } i=\nu-1.\end{cases}$$

The claim is proved.

We now show, by contradiction, that the distributions $\mathscr{G}_i$ must be involutive. Suppose that the distributions $\mathscr{G}_i$ are not all involutive i, $0\leqq i\leqq n-1$. Let $k<n$ be the smallest integer such that $\mathscr{G}_k$ is not involutive; then for some j, $0\leqq j\leqq k-1$,

$$[\mathrm{ad}_f^j g,\mathrm{ad}_f^k g]\notin\mathscr{G}_k. \tag{25}$$

Let l be such that $\gamma_l \neq 0$; according to the claim

$$(26) \qquad [\mathrm{ad}_{\tilde{f}}^{j+\nu}\, \tilde{g}_l, \mathrm{ad}_{\tilde{f}}^{k+\nu}\, \tilde{g}_l] = \left[\begin{pmatrix} \gamma_l \,\mathrm{ad}_f^j\, g + X_{l,j+\nu} \\ * \end{pmatrix}, \begin{pmatrix} \gamma_l \,\mathrm{ad}_f^k\, g + X_{l,k+\nu} \\ * \end{pmatrix}\right],$$

where $X_{l,j+\nu} \in \overline{\mathscr{G}_{j-1}} = \mathscr{G}_{j-1}$ and $X_{l,k+\nu} \in \overline{\mathscr{G}_{k-1}} = \mathscr{G}_{k-1}$: it follows from (25) and (26) that $\tilde{\mathscr{G}}_{k+\nu}$ is not involutive and this contradicts the assumption (ii). We established that $\mathscr{G}_i$ is involutive for every $i, 0 \leqq i \leqq n-1$. We remark that the claim implies that $\mathscr{G}_i = T\pi(\tilde{\mathscr{G}}_{i+\nu})$ where π is the projection $(x, w) \overset{\pi}{\mapsto} x$. To eliminate the input redundancies we can make a suitable choice of basis and decompose $\tilde{\mathscr{G}}_{\nu+i}$ in the following way:

$$\tilde{\mathscr{G}}_{\nu+i} = E_i \oplus F_i,$$

where $T\pi$ induces a bijection from E_i to $\mathscr{G}_i$ and $F_i = \tilde{\mathscr{G}}_{\nu+i} \cap \ker T\pi$. Thus we have

$$\operatorname{rank} \tilde{\mathscr{G}}_{\nu+i} = \operatorname{rank} \mathscr{G}_i + \operatorname{rank} F_i.$$

Since the functions $\tilde{x} \mapsto \operatorname{rank} \mathscr{G}_i(\pi(\tilde{x}))$ and $\tilde{x} \mapsto \operatorname{rank} F_i(\tilde{x})$ are lower semicontinuous and since $\tilde{\mathscr{G}}_i$ has constant rank, F_i and $\mathscr{G}_i$ also have constant rank.

Let $\mathscr{G}_* = \bigcup_{i=0}^{\infty} \mathscr{G}_i$ and $\tilde{\mathscr{G}}_* = \bigcup_{i=0}^{\infty} \tilde{\mathscr{G}}_i$ we have $T\pi(\tilde{\mathscr{G}}_*) = \mathscr{G}_*$. Assumption (ii) implies that $\tilde{\mathscr{G}}_* = T\mathbb{R}^{n+q}$, so we have $\mathscr{G}_* = T\mathbb{R}^n$, but since $\mathscr{G}_* = \mathscr{G}_{n-1}$, $(z \in \mathbb{R}^n)$, we have $\operatorname{rank} \mathscr{G}_{n-1} = n$. □

3. A necessary condition. We give an easy necessary condition for nonlinear system dynamic linearization.

THEOREM 3.1. *If system* (1) *is locally dynamic feedback linearizable, then its linear approximation at the origin* $\dot{z} = \nabla_z f(0) z + G(0) u \triangleq Fz + Gu$ *is controllable, i.e.,* $\operatorname{rank}(G, FG, \cdots, F^{n-1}G) = n$.

Proof. According to the assumptions, there exists a dynamic compensator (9) and an extended state diffeomorphism $x = \varphi(z, w)$ that transform the extended system into a linear controllable one (2). Consider the extended system:

$$(27) \qquad \begin{aligned} \dot{z} &= f(z) + G(z)\alpha(z, w) + G(z)\beta(z, w)v \\ \dot{w} &= a(z, w) + B(z, w)v. \end{aligned}$$

It follows that the linear approximation of (27)

$$(28) \qquad \begin{aligned} \dot{z} &= (\nabla_z f)(0)z + G(0)(\nabla_z \alpha)(0,0)z + G(0)(\nabla_w \alpha)(0,0)w + G(0)\beta(0,0)v \\ \dot{w} &= (\nabla_z a)(0,0)z + (\nabla_w a)(0,0)w + B(0,0)v \end{aligned}$$

is equivalent, up to a linear change of coordinates to a linear controllable system and thus is controllable. A straightforward computation then shows that the controllability of (28) implies that the linear approximation of (1) is controllable, i.e., it satisfies the Kalman criterion. □

This condition is obviously not sufficient, as is shown by the following example:

$$(29) \qquad \begin{aligned} \dot{x}_1 &= x_2 + x_3^2, \\ \dot{x}_2 &= x_3, \\ \dot{x}_3 &= u. \end{aligned}$$

The linear approximation at the origin of this system is

$$\dot{x} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} x + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} u,$$

which is controllable. However, according to Theorem 2.1, the single input system (29) is not static feedback linearizable and, according to Theorem 2.2, it is not dynamic feedback linearizable.

4. Sufficient conditions. In this section we investigate the structure of the distribution $\tilde{\mathcal{G}}_i$ for the following extended system, given by system (1) along with a dynamic compensator (12) or (13)

$$
\begin{aligned}
&\dot{z}=f(z)+\sum_{j,\mu_j\geqq 1} w_1^j g_j(z)+\sum_{j,\mu_j=0} v_j'(t)g_j(z)\\
(30)\qquad &\dot{w}_i^j=w_{i+1}^j \qquad 1\leqq j\leqq m,\quad \mu_j>1,\quad 1\leqq i\leqq \mu_j-1\\
&\dot{w}_{\mu_j}^j=v_j' \qquad 1\leqq j\leqq m,\quad \mu_j\geqq 1,
\end{aligned}
$$

which can be written as

$$
(31)\qquad \dot{\tilde{z}}=\tilde{f}(\tilde{z})+\sum_{j=1}^{m}\tilde{g}_j(\tilde{z})v_j'(t),
$$

where

$$
\begin{aligned}
&\tilde{z}=(z, w_1^j,\cdots,w_{\mu_j}^j;\ 1\leqq j\leqq m,\ \mu_j\geqq 1)=(z,w)\\
&\tilde{f}=f+\sum_{j,\mu_j\geqq 1} w_1^j g_j+\sum_{j,\mu_j\geqq 2}\ \sum_{i=1}^{\mu_j-1} w_{i+1}^j\frac{\partial}{\partial w_i^j}\\
(32)\qquad &\\
&\tilde{g}_j=\begin{cases} g_j & \text{if } \mu_j=0\\ \dfrac{\partial}{\partial w_{\mu_j}^j} & \text{if } \mu_j\geqq 1.\end{cases}
\end{aligned}
$$

The distributions $\tilde{\mathcal{G}}_i$ for system (31) are defined in the extended space (z, w) as

$$
(33)\qquad \tilde{\mathcal{G}}_i=sp\{\mathrm{ad}_{\tilde{f}}^l\,\tilde{g}_j,\ 0\leqq l\leqq i,\ 1\leqq j\leqq m\}.
$$

We now define on the original state space a set of distributions that depend on the indices $\{\mu_1,\cdots,\mu_m\}$ and that play a crucial role in this paper:

$$
\begin{aligned}
&\Delta_0=sp\{g_k,\ 1\leqq k\leqq m,\ \mu_k=0\}\\
(34)\qquad &\\
&\Delta_{i+1}=\Delta_i+\mathrm{ad}_f\Delta_i+sp\{g_k,\ 1\leqq k\leqq m,\ \mu_k=i+1\}.
\end{aligned}
$$

By definition, $\Delta_0\subset\Delta_1\subset\cdots\subset\Delta_i\subset\cdots$. The following lemma gives conditions under which the set of distributions $\tilde{\mathcal{G}}_i$ in the extended state space is clearly related to the set of distributions Δ_i in the original state space.

LEMMA 4.1. *If for all i, $i\geqq 0$ and for all j, $1\leqq j\leqq m$, such that $\mu_j\geqq 1$*

$$
[g_j,\Delta_i]\subset\Delta_{i+1}
$$

then, for $i\geqq 0$

$$
(35)\qquad \tilde{\mathcal{G}}_i=\Delta_i+sp\left\{\frac{\partial}{\partial w_{\mu_j}^j}\,\middle|\,\mu_j\geqq 1\right\}+\cdots+sp\left\{\frac{\partial}{\partial w_{\mu_j-i}^j}\,\middle|\,\mu_j\geqq i+1\right\}.
$$

Proof. Let us compute $\tilde{\mathcal{G}}_0$

$$
\begin{aligned}
\tilde{\mathcal{G}}_0&=sp\{\tilde{g}_1,\cdots,\tilde{g}_m\}\\
&=sp\{g_k\,|\,\mu_k=0\}+sp\left\{\frac{\partial}{\partial w_{\mu_j}^j}\,\middle|\,\mu_j\geqq 1\right\}.
\end{aligned}
$$

By definition of Δ_0 we have:

$$\tilde{\mathscr{G}}_0 = \Delta_0 + sp\left\{\frac{\partial}{\partial w^j_{\mu_j}} \,\middle|\, \mu_j \geqq 1\right\}.$$

We now proceed by induction in the computation of the distributions $\tilde{\mathscr{G}}_i$ by showing that if

$$\tilde{\mathscr{G}}_i = \Delta_i + sp\left\{\frac{\partial}{\partial w^j_{\mu_j}} \,\middle|\, \mu_j \geqq 1\right\} + \cdots + sp\left\{\frac{\partial}{\partial w^j_{\mu_j - i}} \,\middle|\, \mu_j \geqq i+1\right\}$$

then, under the assumption of Lemma 4.1 we have

$$\tilde{\mathscr{G}}_{i+1} = \Delta_{i+1} + sp\left\{\frac{\partial}{\partial w^j_{\mu_j}} \,\middle|\, \mu_j \geqq 1\right\} + \cdots + sp\left\{\frac{\partial}{\partial w^j_{\mu_j - i - 1}} \,\middle|\, \mu_j \geqq i+2\right\},$$

which can be shown by computing

$$\tilde{\mathscr{G}}_{i+1} = \tilde{\mathscr{G}}_i + \mathrm{ad}_{\tilde{f}}\, \tilde{\mathscr{G}}_i.$$

We have

$$\begin{aligned} \mathrm{ad}_{\tilde{f}}\, \tilde{\mathscr{G}}_i &= sp\{\mathrm{ad}_{\tilde{f}}\, X;\ X \in \tilde{\mathscr{G}}_i\} \\ &= sp\left\{\mathrm{ad}_f\, Y + \sum_{j, \mu_j \geqq 1} w^j_1\, \mathrm{ad}_{g_j}\, Y;\ Y \in \Delta_i\right\} + sp\{g_j \mid \mu_j = i+1\} \\ &\quad + sp\left\{\frac{\partial}{\partial w^j_{\mu_j}};\ \mu_j \geqq 1\right\} + \cdots + sp\left\{\frac{\partial}{\partial w^j_{\mu_j - i - 1}};\ \mu_j \geqq i+2\right\}. \end{aligned}$$

Consider $[g_j, \Delta_i]$ for those j such that $\mu_j \geqq 1$; by the assumptions of the lemma we have $[g_j, \Delta_i] \subset \Delta_{i+1}$ for all j such that $\mu_j \geqq 1$. According to the definition of Δ_{i+1}, we have proved that the lemma holds true for $i+1$. Since the induction has been proved for $i = 0$, the lemma holds true for every $i \geqq 0$. □

THEOREM 4.2. *If for a set of integers* $\{\mu_1, \cdots, \mu_m\}$, $0 \leqq \mu_1 \leqq \cdots \leqq \mu_m$, $\mu = \sum_{i=1}^m \mu_i$, *the distributions, up to input reordering,*

$$\Delta_0 = sp\{g_k;\ \mu_k = 0\}$$

$$\Delta_{i+1} = \Delta_i + \mathrm{ad}_f \Delta_i + sp\{g_k;\ \mu_k = i+1\}, \qquad i \geqq 0$$

are such that in U_0, *a neighborhood of the origin*

(i) Δ_i *is involutive and of constant rank for* $0 \leqq i \leqq n + \mu_m - 1$;

(ii) rank $\Delta_{n+\mu_m-1} = n$;

(iii) $[g_j, \Delta_i] \subset \Delta_{i+1}$ *for all* j, $1 \leqq j \leqq m$, *such that* $\mu_j \geqq 1$ *and all* i, $0 \leq i \leq n + \mu_m - 1$; *then the system* (1) *is locally dynamic feedback linearizable by a dynamic compensator* (12) *with indices* $\mu_1, \ldots, \mu_m$ *and a local diffeomorphism in* V_0, *a neighborhood of the origin in the extended state space* $\mathbb{R}^{n+\mu}$.

Proof. According to Lemma 4.1, assumption (iii) guarantees that (35) holds for every $i \geqq 0$. Since Δ_i only depends on z, being defined on the original state space, it follows from (35) and assumption (i) that the distributions $\tilde{\mathscr{G}}_i$ are involutive and of constant rank for every $i \geqq 0$.

By assumption (ii), equality (35) for $i = n + \mu_m - 1$ implies that rank $\tilde{\mathscr{G}}_{n+\mu_m-1} = n + \mu$ and thus rank $\tilde{\mathscr{G}}_{n+\mu-1} = n + \mu$. We can then apply Theorem 2.1 to the extended system (30), which guarantees the existence of an extended state feedback

$$v' = \alpha(z, w) + \beta(z, w) v$$

with $\alpha(0,0)=0, \beta(z,w)$ nonsingular $m\times m$ matrix, and of an extended space diffeomorphism

$$x=\varphi(z,w), \qquad \varphi(0,0)=0 \qquad z\in\mathbb{R}^{n+\mu}$$

transforming the system (30) into a linear controllable one in x-coordinates. In conclusion we have proved that there exist a dynamic state feedback compensator (12) with indices $\mu_1, \cdots \mu_m$ and an extended space diffeomorphism (10) transforming the system (1) into a linear and controllable one of dimension $n+\mu$. □

COROLLARY 4.3. *Consider system* (1) *with* $m=n-1$. *The following statements are equivalent*:

(i) *System* (1) *is locally dynamic feedback linearizable.*

(ii) *The linear approximation of system* (1) *at the origin is controllable.*

Proof. (i)⇒(ii) follows from Theorem 3.1.

(ii)⇒(i). Since the linear approximation of system (1) is controllable, there exists a vector field g_i such that $\mathrm{ad}_f g_i$ does not belong to $\mathscr{G}_0$; furthermore $\mathscr{G}_1=T\mathbb{R}^n$. There are two cases. Either $\mathscr{G}_0$ is involutive and therefore, since $\mathscr{G}_1=T\mathbb{R}^n$, according to Theorem 2.1, system (1) is static feedback linearizable. In the opposite case, we set $\mu_k=1$ if $k\neq i$ and $\mu_i=0$, we have

- $\Delta_0=sp\{g_i\}$ which is of rank 1 and involutive.
- $\Delta_1=\mathscr{G}_0+sp\{\mathrm{ad}_f g_i\}=T\mathbb{R}^n$ since $n=m+1$.

The assumptions of Theorem 4.2 are met and system (1) is dynamic feedback linearizable. □

Remark 1. Corollary 4.3 agrees with the sufficient conditions obtained in [5]. Corollary 4.3 is false if $m<n-1$ in system (1), as shown by (29) (single-input system). A multi-input counterexample is given by the following system

$$\dot{x}_1=x_2+x_3^2$$

$$\dot{x}_2=x_3$$

$$\dot{x}_3=u_1$$

$$\dot{x}_4=u_2$$

Its linear approximation at the origin is controllable; nevertheless, it can be easily seen that the system is a dynamic extension of (29) by setting $m'=2$, $w_1=x_4$, $\dot{w}_1=u_2$, and $u=u_1$ with the notations of (14). Since (29) is not dynamic feedback linearizable by virtue of Theorem 2.2, it is also true for any dynamic extension of (29) (by contradiction), which proves that our multi-input system is not dynamic feedback linearizable.

Remark 2. Conditions (i) and (iii) of Theorem 4.2 have some redundancy. In fact the involutivity of Δ_{i+1} and the inclusion $\Delta_i\subset\Delta_{i+1}$ imply

$$[g_j, \Delta_{i+1}]\subset\Delta_{i+1}$$

and therefore

$$[g_j, \Delta_i]\subset\Delta_{i+1}$$

for all j, $1\leqq j\leqq m$, such that $\mu_j\leqq i+1$. Hence only the conditions

$$[g_j, \Delta_i]\subset\Delta_{i+1}$$

for all $i\geqq 0$ and for all j, $1\leqq j\leqq m$, such that $\mu_j>i+1$ are not contained in assumption (i).

Remark 3. The conditions of Theorem 4.2 are invariant under state space diffeomorphism (3), but they are not invariant under feedback transformations (5).

Remark 4. If $\mu_i = 0, 1 \leqq i \leqq m$, then $\Delta_j = \mathscr{G}_j$ for every $j \geqq 0$. Since condition (i) of Theorem 4.2 requires that the distributions $\Delta_j = \mathscr{G}_j$ be involutive, condition (iii) of Theorem 4.2 is always satisfied (in fact $[g_j, \Delta_i] \subset \Delta_i$, $1 \leqq j \leqq m$, in this case). Conditions (i) and (ii) of Theorem 4.2 coincide then with the necessary and sufficient conditions of Theorem 2.1. Theorem 4.2 can be viewed as a generalization of Theorem 2.1.

Remark 5. For those systems for which Theorem 4.2 applies, we can compute the controllability indices of system (30) from (35) on the basis of the distributions Δ_i. Let

$$\begin{aligned} \tilde{m}_0 &= \text{rank } \tilde{\mathscr{G}}_0 = m \\ \tilde{m}_i &= \text{rank } \tilde{\mathscr{G}}_i - \text{rank } \tilde{\mathscr{G}}_{i-1} \\ &= \text{rank } \Delta_i - \text{rank } \Delta_{i-1} + \text{card } \{j \mid \mu_j \geqq i+1;\ 1 \leqq j \leqq m\}, \end{aligned}$$

then

$$\tilde{k}_j = \text{card}\,\{i \mid \tilde{m}_i \geqq j\}, \qquad 1 \leqq j \leqq m,$$

is such that $\sum_{j=1}^{m} \tilde{k}_j = n + \mu$. Up to an input renumbering, we can then associate to any vector field g_j an index $\tilde{k}_j - \mu_j$ so that

$$\text{rank}\,\{g_j, \cdots, \text{ad}_f^{\tilde{k}_j - \mu_j - 1} g_j;\ 1 \leqq j \leqq m\} = n.$$

System (30) can be transformed into a linear controllable system with controllability indices $(\tilde{k}_1, \cdots, \tilde{k}_m)$ as follows:

1. Determine m functions $\varphi_1(z), \cdots, \varphi_m(z)$ such that:

   (i) The $m \times m$ matrix

   $$D(x) = (\langle \text{d}\varphi_i, \text{ad}_f^{\tilde{k}_i - \mu_j - 1} g_j \rangle)$$

   is nonsingular;

   (ii) $\langle \text{d}\varphi_i, \text{ad}_f^{\lambda} g_j \rangle = 0$, for every j, $1 \leqq j \leqq m$, and every λ, $0 \leqq \lambda < \tilde{k}_i - \mu_j - 1$.

2. The extended state space diffeomorphism is given by

   $$\varphi(z, w) = (\varphi_1, \cdots, L_{\tilde{f}}^{\tilde{k}_1 - 1} \varphi_1, \cdots, \varphi_m, \cdots, L_{\tilde{f}}^{\tilde{k}_m - 1} \varphi_m)$$

   with $\tilde{f}$ and $\tilde{g}_i$ being given by (30) and (32).

3. The extended state feedback transformation is (up to the former input renumbering)

   $$v'(t) = \alpha(z, w) + \beta(z, w) v(t) = \tilde{D}^{-1} \left[\begin{pmatrix} L_{\tilde{f}}^{\tilde{k}_1} \varphi_1(z, w) \\ \vdots \\ L_{\tilde{f}}^{\tilde{k}_m} \varphi_m(z, w) \end{pmatrix} + v(t) \right]$$

   with

   $$\tilde{D}_j^i(z, w) = L_{\tilde{g}_i} L_{\tilde{f}}^{\tilde{k}_j - 1} \varphi_j(z, w).$$

Remark 6. The sufficient conditions of Theorem 4.2 may be helpful in finding the structure of a dynamic compensator if it exists. We sketch the analysis in a particular case.

Assume that $\Lambda_0 = sp\{g_1, \cdots, g_{m-1}\}$ is an involutive subdistribution of $\mathscr{G}_0$. We assume that for each $i \geqq 0$, we have

$$\overline{\mathscr{G}_i} = \mathscr{G}_i + I(\{\text{ad}_f^k g_m, 0 \leqq k \leqq i\})$$

where $I(\{\text{ad}_f^k g_m, 0 \leqq k \leqq i\})$ is the Lie ideal of the Lie algebra $\overline{\mathscr{G}_i}$ generated by $\{\text{ad}_f^k g_m, 0 \leqq k \leqq i\}$. These conditions mean that the noninvolutivity of the distributions $\mathscr{G}_i$ is due to brackets involving g_m.

This specific Lie algebra structure of the distributions $\mathcal{G}_i$ and $\overline{\mathcal{G}_i}$ leads to the construction of a dynamic compensator of type (12) with $\mu_1=\cdots=\mu_{m-1}=0$ and $\mu_m\neq 0$. It remains to choose the value of μ_m. The following procedure gives bounds on the length of the chain of integrators.

We define the following distributions and indices

$$\Lambda_{i+1}=\Lambda_i+\mathrm{ad}_f\,\Lambda_i$$

$$i_0=\inf\{i\,|\,\mathcal{G}_i\neq\overline{\mathcal{G}_i}\}$$

$$\mu=\sup\{0,\{i_0-k\,|\,[\mathrm{ad}_f^k\,g_m,\mathrm{ad}_f^{i_0}\,\Lambda_0]\not\subset\mathcal{G}_{i_0},0\leqq k\leqq i_0\}\}$$

$$r=\sup\{i\,|\,\Lambda_i \text{ involutive and } \mathrm{ad}_{g_m}\,\Lambda_{i-1}\subset\Lambda_i\}.$$

Let $\mu_1=\cdots=\mu_{m-1}=0$. Assume that $\mu\leqq r$. Then let us choose μ_m and the distributions Δ_i as follows:

$$(36)\qquad\begin{aligned}&\mu+1\leqq\mu_m\leqq r+1\\&\Delta_i=\Lambda_i && \forall i\leqq\mu_m-1\\&\Delta_{\mu_m}=\Lambda_{\mu_m}+sp\{g_m\}\\&\Delta_{i+1}=\Delta_i+\mathrm{ad}_f\Delta_i && \forall i\geq\mu_m.\end{aligned}$$

According to the definition of the distributions Λ_i and of the index r, to Theorem 4.2 and to Remark 2, if Δ_i is involutive for all $i\geqq\mu_m$ then system (1) is dynamic feedback linearizable by the dynamic compensator (12) defined with $\mu_1=\cdots=\mu_{m-1}=0$ and μ_m given by (36).

If μ_m is chosen less than or equal to μ, Δ_{i_0} cannot be involutive:

$$\Delta_{i_0}=\sum_{k=0}^{i_0}\mathrm{ad}_f^k\,\Lambda_0+sp\{\mathrm{ad}_f^k\,g_m,0\leqq k\leqq i_0-\mu_m\},$$

$\mathrm{ad}_f^{i_0-\mu}\,g_m\subset\Delta_{i_0}$ which implies, according to the definition of μ, that $\Delta_{i_0}\subset\mathcal{G}_{i_0}$ is not involutive.

On the contrary, if the index μ_m is chosen strictly greater than $r+1$, then

$$\Delta_{r+1}=\Lambda_{r+1}$$

and according to the definition of r, Δ_{r+1} is not involutive. Therefore, the only possible choice is $\mu+1\leqq\mu_m\leqq r+1$.

5. Examples.

Example 1. This example illustrates how the sufficient conditions of Theorem 4.2 can lead to the construction of a linearizing dynamic compensator. Consider the system

$$\begin{aligned}\dot x_1&=x_2,\\\dot x_2&=u_1,\\\dot x_3&=u_2,\\\dot x_4&=x_3-x_3u_1;\end{aligned}$$

that is,

$$\dot x=f(x)+u_1g_1(x)+u_2g_2(x)$$

with

$$f = x_2 \frac{\partial}{\partial x_1} + x_3 \frac{\partial}{\partial x_4},$$

$$g_1 = \frac{\partial}{\partial x_2} - x_3 \frac{\partial}{\partial x_4},$$

$$g_2 = \frac{\partial}{\partial x_3}.$$

Easy computations give

$$\mathrm{ad}_{g_1} g_2 = \frac{\partial}{\partial x_4},$$

$$\mathrm{ad}_f g_2 = -\frac{\partial}{\partial x_4},$$

$$\mathrm{ad}_f g_1 = -\frac{\partial}{\partial x_1}.$$

The system is not feedback linearizable since $\mathscr{G}_0$ is not involutive and, therefore, condition (i) of Theorem 2.1 fails, whereas condition (ii) is satisfied, i.e., rank $\mathscr{G}_1 = 4$. Analyzing the Lie algebraic structure of the system we have that $sp\{g_1, g_2, \mathrm{ad}_f g_2\}$ is involutive and of constant rank 3 whereas $sp\{g_1, g_2, \mathrm{ad}_f g_1\}$ is not involutive. According to Remark 6, this leads us to choose $\mu_1 = 1$ and $\mu_2 = 0$, so that we have

$$\Delta_0 = sp\{g_2\},$$

$$\Delta_1 = sp\{g_2, \mathrm{ad}_f g_2, g_1\},$$

$$\Delta_2 = sp\{g_2, g_1, \mathrm{ad}_f g_1, g_2\}.$$

The conditions of Theorem 4.2 are satisfied. The controllability indices are $(3, 2)$ and

$$\mathrm{rank}\,\{g_1, \mathrm{ad}_f g_1, g_2, \mathrm{ad}_f g_2\} = 4.$$

The dynamic compensator is then

$$\frac{\mathrm{d}u_1}{\mathrm{d}t} = v_1', \qquad u_2 = v_2'.$$

Following the procedure outlined in Remark 5, the extended space change of coordinates is obtained solving

$$\mathrm{d}\varphi_1 \in (\Delta_1)^\perp, \qquad \mathrm{d}\varphi_2 \in (\Delta_0)^\perp$$

with

$$\begin{pmatrix} \langle \mathrm{d}\varphi_1, \mathrm{ad}_f g_1 \rangle & \langle \mathrm{d}\varphi_1, \mathrm{ad}_f^2 g_2 \rangle \\ \langle \mathrm{d}\varphi_2, g_1 \rangle & \langle \mathrm{d}\varphi_1, \mathrm{ad}_f g_2 \rangle \end{pmatrix}$$

nonsingular; a solution is $\varphi_1 = x_1$, $\varphi_2 = x_4$. The diffeomorphism is then

$$\varphi = (\varphi_1, L_{\tilde f}\varphi_1, L_{\tilde f}^2\varphi_1, \varphi_2, L_{\tilde f}\varphi_2) = (x_1, x_2, u_1, x_4, x_3 - x_3 u_1)$$

and the extended state feedback transformation is

$$\begin{pmatrix} v_1' \\ v_2' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -x_3 & 1 - u_1 \end{pmatrix}^{-1} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}.$$

The dynamic compensation has a singularity at $u_1 = 1$ where there is a loss of controllability.

Example 2. This example shows that condition (iii) of Theorem 4.2 is not necessary. Consider the system

$$\begin{aligned} \dot{x}_1 &= x_2 + x_3 u_2, \\ \dot{x}_2 &= x_3 + x_1 u_2, \\ \dot{x}_3 &= u_1 + x_2 u_2, \\ \dot{x}_4 &= u_2; \end{aligned}$$

i.e.,

$$\dot{x} = f(x) + u_1 g_1(x) + u_2 g_2(x)$$

with

$$f = x_2 \frac{\partial}{\partial x_1} + x_3 \frac{\partial}{\partial x_2}, \quad g_1 = \frac{\partial}{\partial x_3}, \quad g_2 = x_3 \frac{\partial}{\partial x_1} + x_1 \frac{\partial}{\partial x_2} + x_2 \frac{\partial}{\partial x_3} + \frac{\partial}{\partial x_4}.$$

We can compute

$$\mathrm{ad}_{g_2} g_1 = -\frac{\partial}{\partial x_1}, \quad \mathrm{ad}_{g_2}^2 g_1 = \frac{\partial}{\partial x_2}, \quad \mathrm{ad}_{g_2}^3 g_1 = -\frac{\partial}{\partial x_3};$$

that is, rank $\overline{\mathscr{G}_0} = 4$, and

$$\mathrm{ad}_f g_1 = -\frac{\partial}{\partial x_2}, \quad \mathrm{ad}_f^2 g_1 = \frac{\partial}{\partial x_1}, \quad \mathrm{ad}_f g_2 = -x_1 \frac{\partial}{\partial x_1} + x_3 \frac{\partial}{\partial x_3}.$$

Since $\mathscr{G}_0$ is not involutive, the system is not static feedback linearizable; however, rank $\mathscr{G}_2 = 4$ and the linear approximation at the origin is controllable. If we set $\mu_1 = 0$ and $\mu_2 = 3$, the extended system

$$\begin{aligned} \dot{x}_1 &= x_2 + x_3 w_1, \\ \dot{x}_2 &= x_3 + x_1 w_1, \\ \dot{x}_3 &= v_1' + x_2 w_1, \\ \dot{x}_4 &= w_1, \\ \dot{w}_1 &= w_2, \\ \dot{w}_2 &= w_3, \\ \dot{w}_3 &= v_2' \end{aligned}$$

turns out to be static feedback linearizable; in fact,

$$\tilde{\mathscr{G}}_0 = sp\left\{\frac{\partial}{\partial w_3}, \frac{\partial}{\partial x_3}\right\}$$

$$\tilde{\mathscr{G}}_1 = \tilde{\mathscr{G}}_0 + sp\left\{\frac{\partial}{\partial w_2}, \frac{\partial}{\partial x_2} + w_1 \frac{\partial}{\partial x_1}\right\}$$

$$\tilde{\mathscr{G}}_2 = \tilde{\mathscr{G}}_1 + sp\left\{\frac{\partial}{\partial w_1}, (1 - w_2)\frac{\partial}{\partial x_1} + w_1^2 \frac{\partial}{\partial x_2} + w_1 \frac{\partial}{\partial x_3}\right\}.$$

When $w_1^3+w_2-1\neq 0$ (i.e., when u_1 satisfies $\dot{u}_1\neq 1-u_1^3$) we have

$$\tilde{G}_2=\tilde{G}_1+sp\left\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}\right\}.$$

Thus rank $\tilde{\mathscr{G}}_2=6$ and, again when $w_1^3+w_2-1\neq 0$, we have $\tilde{\mathscr{G}}_3=T\mathbb{R}^7$. Hence the distributions $\tilde{\mathscr{G}}_i$ satisfy the conditions of Theorem 2.1 as long as $w_1^3+w_2-1\neq 0$. On the other hand, $\Delta_0=sp\{\partial/\partial x_3\}$, $\Delta_1=sp\{(\partial/\partial x_3), (\partial/\partial x_2)\}$ and condition (iii) of Theorem 4.2 fails since $\mathrm{ad}_{g_2} g_1$ is not contained in Δ_1; note that formula (35) does not hold for $i=1$.

This example illustrates the role of condition (iii): although not necessary, it allows us to check the involutivities and constant rank condition for the distributions $\tilde{\mathscr{G}}_i$ on the basis of the distributions Δ_i defined on the original state space.

Example 3. Theorem 6.5 of [12, p. 136] was the first general result where dynamic feedback linearization was achieved. The present example shows that the sufficient conditions obtained in Theorem 4.2 and Corollary 4.3 are different than those given in [12], which we now recall. The notations are those of [12]. Define the codistributions Ω_k, $k\geqq 0$:

$$\Omega_0=sp\{\mathrm{d}h\}=sp\{\mathrm{d}h_1,\cdots,\mathrm{d}h_p\},$$

$$\Omega_{k+1}=\Omega_k+L_f(\Omega_k\cap\mathscr{G}_0^\perp)+\sum_i L_{g_i}(\Omega_k\cap\mathscr{G}_0^\perp),$$

$$\Omega^*=\bigcap_{k=0}^{\infty}\Omega_k.$$

Recall that if $\omega=\sum_i \omega_i\,dx_i$ is a smooth 1-form and $\gamma=\sum_i \gamma_i(\partial/\partial x_i)$ is a smooth vector field, we have

$$L_\gamma\omega=\left(\frac{\partial\omega^T}{\partial x}\gamma\right)^T+\omega\frac{\partial\gamma}{\partial x}.$$

Recall also that $\Delta^*=(\Omega^*)^\perp$ is the largest (f,g)-invariant distribution contained in $\mathrm{Ker}\,\mathrm{d}h=\Omega_0^\perp$.

In [12] the following result is proved.

Assume that:

(A1) $$\Delta^*=\{0\},$$

(A2) $$\sum_i L_{g_i}(\Omega_k\cap\mathscr{G}_0^\perp)\subset\Omega_k \textit{ for every } k\geqq 0.$$

Then the system is locally dynamic feedback linearizable.

Consider the system, which was suggested to us by Respondek:

$$\begin{aligned}\dot{x}_1&=u_1,\\ \dot{x}_2&=x_3+\varphi(x)u_1,\\ \dot{x}_3&=u_2,\\ y_1&=h_1(x),\\ y_2&=h_2(x),\end{aligned}\tag{37}$$

with $\varphi(0)=0$, $(\partial\varphi/\partial x_3)(x)\neq 0$, for all $x\in\mathbb{R}^3$, and h_1, h_2 two generic output functions. The linear approximation at the origin is controllable and Corollary 4.3 applies. Thus, h_1, h_2 can be constructed according to Remark 5.

Let us prove that there cannot exist two analytic output functions h_1, h_2 for the system (37) such that the sufficient conditions of [12] are satisfied.

We have

$$f = x_3 \frac{\partial}{\partial x_2}$$

and

$$g_1 = \frac{\partial}{\partial x_1} + \varphi \frac{\partial}{\partial x_2}, \qquad g_2 = \frac{\partial}{\partial x_3}.$$

Thus $\mathscr{G}_0^{\perp} = sp\{(\varphi, -1, 0)\}$. Since $\mathscr{G}_0^{\perp}$ is one-dimensional, two cases can arise:

- $\mathscr{G}_0^{\perp} \cap \Omega_0 = 0$, which implies that $\Omega^* = \Omega_0$ and thus Δ^* is one-dimensional, $\Delta^* \neq 0$. Thus (A1) is violated.
- $\mathscr{G}_0^{\perp} \cap \Omega_0 = \mathscr{G}_0^{\perp}$ and we have $L_{g_1}(\varphi, -1, 0) = (\varphi(\partial\varphi/\partial x_2), -(\partial\varphi/\partial x_2), -(\partial\varphi/\partial x_3))$, $L_{g_2}(\varphi, -1, 0) = ((\partial\varphi/\partial x_3), 0, 0)$. It results that, if $(\partial\varphi/\partial x_3) \neq 0$, $\Omega_0 + \sum_i L_{g_i}(\Omega_0 \cap \mathscr{G}_0^{\perp}) = T^*\mathbb{R}^3 \neq \Omega_0$, which proves that (A2) is violated.

The claim is proved.

Example 4 (rigid body dynamics). We consider a general model of rigid body dynamics that includes the case of aircraft dynamics [5]. Our general purpose here is to show that Theorem 4.2 applies to important classes of mechanical systems rather than deal with specific applications of aircraft control.

Let (x, y, z) be the coordinates of the center of mass in an absolute frame with the vertical z-axis oriented downward, (u, v, w) the velocity components in a relative frame linked to the rigid body, (p, q, r) the components of the kinetic moment in the relative frame, (Φ, Θ, Ψ) the roll, pitch, and yaw angles, respectively. Let $\xi = (x, y, z, u, v, w, \Phi, \Theta, \Psi)^T$ be the state vector.

We consider p, q, r, and ρ (ρ is the thrust) as control variables. The equations of motion are

$$\begin{aligned}
\dot{x} &= u \cos\Psi \cos\Theta + v(\cos\Psi \sin\Theta \sin\Phi - \sin\Psi \cos\Phi) \\
&\quad + w(\cos\Psi \sin\Theta \cos\Phi + \sin\Psi \sin\Phi), \\
\dot{y} &= u \sin\Psi \cos\Theta + v(\sin\Psi \sin\Theta \sin\Phi + \cos\Psi \cos\Phi) \\
&\quad + w(\sin\Psi \sin\Theta \cos\Phi - \cos\Psi \sin\Phi), \\
\dot{z} &= -u \sin\Theta + v \cos\Theta \sin\Phi + w \cos\Theta \cos\Phi, \\
\dot{u} &= -g \sin\Theta + rv - qw + \frac{X_0(\xi)}{m} + \frac{J\rho}{m}, \qquad (38) \\
\dot{v} &= g \cos\Theta \sin\Phi - ru + pw + \frac{Y(\xi)}{m}, \\
\dot{w} &= g \cos\Theta \cos\Phi + qu - pv + \frac{Z(\xi)}{m}, \\
\dot{\Phi} &= p + \mathrm{tg}\Theta(q \sin\Phi + r \cos\Phi), \\
\dot{\Theta} &= q \cos\Phi - r \sin\Phi, \\
\dot{\Psi} &= \frac{q \sin\Phi + r \cos\Phi}{\cos\Theta},
\end{aligned}$$

where $(X_0(\xi)+J\rho,\ Y(\xi),\ Z(\xi))$ are the components of the force vector excepting gravity. The model is of the type

$$\dot{\xi}=f(\xi)+pg_1(\xi)+qg_2(\xi)+rg_3(\xi)+\rho g_4(\xi) \tag{39}$$

with f, g_1, g_2, g_3, and g_4 obtained from (38).

The reader can compute the Lie brackets $\mathrm{ad}_{g_i} g_j$ with $1\leqq i,j\leqq 4$ and check that, independently of the type of functions that $X_0(\xi)$, $Y(\xi)$, and $Z(\xi)$ may be in specific cases,

- $sp\{g_1, g_2, g_3\}$ is involutive;
- $\mathscr{G}_0=sp\{g_1, g_2, g_3, g_4\}$ is not involutive, rank $\mathscr{G}_0=4$;
- $\overline{\mathscr{G}_0}=\mathscr{G}_0+sp\{\mathrm{ad}_{g_4} g_2, \mathrm{ad}_{g_4} g_3\}$ $=sp\{(\partial/\partial u), (\partial/\partial v), (\partial/\partial w), (\partial/\partial\Phi), (\partial/\partial\Theta), (\partial/\partial\Psi)\}$, rank $\overline{\mathscr{G}_0}=6$;
- $\mathscr{G}_1=\mathscr{G}_0+\mathrm{ad}_f\mathscr{G}_0\neq\overline{\mathscr{G}_1}=T\mathbb{R}^9$, rank $\mathscr{G}_1=7$ and rank $\overline{\mathscr{G}_1}=9$.

Since $\mathscr{G}_0$ is not involutive, the system is not feedback linearizable. However, it is clear from the Lie bracket configuration that all the noninvolutivities are caused by g_4 and, according to Remark 6, if we choose $\mu_1=\mu_2=\mu_3=0$ and $\mu_4=1$, we have

$$\Delta_0=sp\{g_1, g_2, g_3\}$$
$$\Delta_1=\Delta_0+\mathrm{ad}_f\,\Delta_0+sp\{g_4\}=\overline{\mathscr{G}_0}$$
$$\Delta_2=\Delta_1+\mathrm{ad}_f\,\Delta_1=T\mathbb{R}^9.$$

Therefore, the sufficient conditions of Theorem 4.2 are satisfied, which leads to the compensator

$$\begin{aligned}\dot{\rho}&=v_4,\\ p&=v_1,\\ q&=v_2,\\ r&=v_3.\end{aligned} \tag{40}$$

Define the functions

$$\begin{aligned}y_1&=x,\\ y_2&=y,\\ y_3&=z,\\ y_4&=\rho,\end{aligned}$$

and the extended space diffeomorphism

$$(\xi,\rho)\mapsto\tilde{\xi},$$

where

$$\tilde{\xi}=(y_1, L_{\tilde{f}}y_1, L_{\tilde{f}}^2y_1, y_2, L_{\tilde{f}}y_2, L_{\tilde{f}}^2y_2, y_3, L_{\tilde{f}}y_3, L_{\tilde{f}}^2y_3, y_4)^T.$$

The dynamic compensator (40) together with extended state feedback

$$v=D^{-1}(v'-D_0),$$

where

$$D_0^1 = \frac{L_f X_0}{m} \cos \Psi \cos \Theta + \frac{L_f Y}{m} (\cos \Psi \sin \Theta \sin \Phi - \sin \Psi \cos \Phi)$$
$$+ \frac{L_f Z}{m} (\cos \Psi \sin \Theta \cos \Phi + \sin \Psi \sin \Phi),$$

$$D_0^2 = \frac{L_f X_0}{m} \sin \Psi \cos \Theta + \frac{L_f Y}{m} (\sin \Psi \sin \Theta \sin \Phi + \cos \Psi \cos \Phi)$$
$$+ \frac{L_f Z}{m} (\sin \Psi \sin \Theta \cos \Phi - \cos \Psi \sin \Phi),$$

$$D_0^3 = \frac{L_f X_0}{m} \sin \Theta + \frac{L_f Y}{m} \cos \Theta \sin \Phi + \frac{L_f Z}{m} \cos \Theta \cos \Phi,$$

$$D_0^4 = 0,$$

and

$$D_1^1 = \frac{Y}{m} (\cos \Psi \sin \Theta \cos \Phi) - \frac{Z}{m} (\cos \Psi \sin \Theta \sin \Phi) + \varepsilon_1^1,$$

$$D_1^2 = -\frac{X_0 + J\rho}{m} (\cos \Psi \sin \Theta \cos \Phi + \sin \Psi \sin \Phi) + \frac{Z}{m} \cos \Psi \cos \Theta + \varepsilon_1^2,$$

$$D_1^3 = \frac{X_0 + J\rho}{m} (\cos \Psi \sin \Theta \sin \Phi - \sin \Psi \cos \Phi) - \frac{Y}{m} \cos \Psi \cos \Theta + \varepsilon_1^3,$$

$$D_1^4 = \frac{J}{m} \cos \Psi \cos \Theta,$$

$$D_2^1 = \frac{Y}{m} (\sin \Psi \sin \Theta \cos \Phi - \cos \Psi \sin \Phi) - \frac{Z}{m} (\sin \Psi \sin \Theta \sin \Phi + \cos \Psi \cos \Phi) + \varepsilon_2^1,$$

$$D_2^2 = -\frac{X_0 + J\rho}{m} (\sin \Psi \sin \Theta \cos \Phi - \cos \Psi \sin \Phi) + \frac{Z}{m} \sin \Psi \cos \Theta + \varepsilon_2^2,$$

$$D_2^3 = \frac{X_0 + J\rho}{m} (\sin \Psi \sin \Theta \sin \Phi + \cos \Psi \cos \Phi) - \frac{Y}{m} \sin \Psi \cos \Theta + \varepsilon_2^3,$$

$$D_2^4 = \frac{J}{m} \sin \Psi \cos \Theta,$$

$$D_3^1 = \frac{\cos \Theta}{m} (-Y \cos \Phi + Z \sin \Phi) + \varepsilon_3^1,$$

$$D_3^2 = \frac{X_0 + J\rho}{m} \cos \Theta \cos \Phi + \frac{Z}{m} \sin \Theta + \varepsilon_3^2,$$

$$D_3^3 = -\left(\frac{X_0 + J\rho}{m} \cos \Theta \sin \Phi + \frac{Y}{m} \sin \Theta \right) + \varepsilon_3^3,$$

$$D_3^4 = \frac{J}{m} \sin \Theta,$$

and $D_4^1 = D_4^2 = D_4^3 = 0$ and $D_4^4 = 1$, with

$$\varepsilon_1^i = \cos\Psi\cos\Theta\,\frac{L_{g_i}X_0}{m} + (\cos\Psi\sin\Theta\sin\Phi - \sin\Psi\cos\Phi)\frac{L_{g_i}Y}{m}$$

$$+ (\cos\Psi\sin\Theta\cos\Phi + \sin\Psi\sin\Phi)\frac{L_{g_i}Z}{m} \qquad 1 \leqq i \leqq 3,$$

$$\varepsilon_2^i = \sin\Psi\cos\Theta\,\frac{L_{g_i}X_0}{m} + (\sin\Psi\sin\Theta\sin\Phi + \cos\Psi\cos\Phi)\,\frac{L_{g_i}Y}{m}$$

$$+ (\sin\Psi\sin\Theta\cos\Phi - \cos\Psi\sin\Phi)\frac{L_{g_i}Z}{m} \qquad 1 \leqq i \leqq 3,$$

$$\varepsilon_3^i = \sin\Theta\,\frac{L_{g_i}X_0}{m} + \cos\Theta\sin\Phi\,\frac{L_{g_i}Y}{m} + \cos\Theta\cos\Phi\,\frac{L_{g_i}Z}{m} \qquad 1 \leqq i \leqq 3$$

makes the closed loop system linear and controllable:

$$\frac{d^3 y_i}{dt^3} = v_i', \qquad 1 \leqq i \leqq 3$$

$$\dot{y}_4 = v_4'.$$

6. Conclusions. We have studied the problem of dynamic feedback linearization (Definition 2.1 and Definition 2.3) for nonlinear systems (1). Theorem 2.2 shows that for single input systems the set of dynamic feedback linearizable systems coincides with the set of static feedback linearizable systems. Theorem 3.1 shows that the controllability of the linear approximation at the origin is a necessary condition for dynamic feedback linearization, which is also sufficient when $m = n - 1$ in (1); this means that mild conditions identify dynamic feedback linearizable systems when the number of states minus the number of controls is equal to one. This is no longer true if $m < n - 1$.

Theorem 4.2 gives sufficient conditions for dynamic feedback linearization via special types of dynamic compensators (prolongations). The conditions are not necessary (Example 2) but are more general than existing sufficient conditions (Example 3) and include the known necessary and sufficient conditions for static feedback linearization. A contrived example (Example 1) and a model of complex dynamical system, namely the dynamics of a rigid body (Example 4), show how the sufficient conditions of Theorem 4.2 lead to the explicit determination of the linearizing dynamic compensator.

The challenging problem of finding necessary and sufficient conditions for dynamic feedback linearization remains open however. We have solved the problem only when $m = n - 1$ and we have established that it is a multi-input phenomenon; in the general case we have obtained constructive sufficient (but not necessary) conditions and necessary conditions.

REFERENCES

[1] M. D. Di Benedetto, J. W. Grizzle, and C. H. Moog, *Rank invariants of nonlinear systems*, SIAM J. Control Optim., 27 (1989), pp. 658–672.

[2] M. D. Di Benedetto and A. Isidori, *The matching of nonlinear models via dynamic state feedback*, SIAM J. Control Optim., 24 (1986), pp. 1063–1075.

[3] R. W. Brockett, *Feedback invariants for nonlinear systems*, Proc. VII IFAC Congress, Helsinki, 1978, pp. 1115-1120.

[4] B. Charlet, J. Lévine, and R. Marino, *Two sufficient conditions for dynamic feedback linearization*, in Analysis and Optimization of Systems, Lecture Notes in Control and Information Sci., Vol. 111, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, Berlin, New York, (1988) pp. 181-192.

[5] ———, *On dynamic feedback linearization*, Systems Control Lett., 13 (1989), pp. 143-151.

[6] ———, *New sufficient conditions for dynamic feedback linearization*, Proceedings of the IFAC Symposium on Nonlinear Control Systems Design, Capri, June 14-16, 1989.

[7] J. Descusse and C. H. Moog, *Decoupling with dynamic compensation for strong invertible affine nonlinear systems*, Internat. J. Control, 42 (1985), pp. 1387-1398.

[8] ———, *Dynamic decoupling for right invertible nonlinear systems*, Systems Control Lett., 8 (1987), pp. 345-349.

[9] M. Fliess, *Généralisation non linéaire de la forme canonique de commande et linéarisation par bouclage*, C.R. Acad. Sci. Paris, Série I, 308 (1989), pp. 377-379.

[10] ———, *Automatique et corps différentiels*, Forum Math., 1 (1989), pp. 227-238.

[11] L. R. Hunt, R. Su, and G. Meyer, *Design for multi-input nonlinear systems*, in Differential Geometric Control Theory, R. Brockett, R. Millman, and H. Sussmann, eds., Birkhäuser, Basel, 1983, pp. 268-298.

[12] A. Isidori, *Control of nonlinear systems via dynamic state feedback*, in Algebraic and Geometric Methods in Nonlinear Control Theory, M. Hazewinkel and M. Fliess, eds., D. Reidel, Boston, 1986, pp. 121-145.

[13] ———, *Nonlinear control systems*, Second edition, Communications and Control Engineering Series, Springer-Verlag, Berlin, New York, 1989.

[14] A. Isidori, C. H. Moog, and A. De Luca, *A sufficient condition for full linearization via dynamic state feedback*, Proc. 25th IEEE CDC, Athens, 1986, pp. 203-207.

[15] B. Jakubczyk and W. Respondek, *On linearization of control systems*, Bull. Acad. Pol. Sci., Ser. Sci. Math., 28 (1980), pp. 517-522.

[16] A. J. Krener, *On the equivalence of control systems and the linearization of nonlinear systems*, SIAM J. Control, 11 (1973), pp. 670-676.

[17] A. J. Krener, A. Isidori, and W. Respondek, *Partial and robust linearization by feedback*, Proc. 22nd IEEE CDC, San Antonio, Texas, 1983, pp. 126-130.

[18] R. Marino, *On the largest feedback linearizable subsystem*, Systems Control Lett., 6 (1986), pp. 345-351.

[19] A. S. Morse, *Structural invariance of linear multivariable systems*, SIAM J. Control, 11 (1973), pp. 446-465.

[20] A. S. Morse and W. M. Wonham, *Decoupling and pole assignment by dynamic compensation*, SIAM J. Control, 8 (1970), pp. 317-337.

[21] H. Nijmeijer and W. Respondek, *Dynamic input-output decoupling of nonlinear control systems*, IEEE Trans. Autom. Control, AC-33 (1988), pp. 1065-1070.

[22] H. Nijmeijer and A. J. van der Schaft, *Controlled invariance for nonlinear systems*, IEEE Trans. Automat. Control, AC-27 (1982), pp. 904-914.

[23] S. Singh, *Decoupling of invertible nonlinear systems with state feedback and precompensation*, IEEE Trans. Automat. Control, AC-25 (1980), pp. 1237-1239.

[24] H. J. Sussmann, *Lie brackets, real analyticity and geometric control*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millmann, and H. J. Sussmann, eds., Birkhäuser, Boston, MA, 1983, pp. 1-116.

[25] D. G. Taylor, P. V. Kokotovic, R. Marino, and I. Kanellakopoulos, *Adaptive regulation of nonlinear systems with unmodelled dynamics*, IEEE Trans. Automat. Control, AC-34 (1989), pp. 405-412.

[26] A. J. van der Schaft, *Linearization and input-output decoupling for general nonlinear systems*, Systems Control Lett., 5 (1984), pp. 27-33.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 1, pp. 58-70, January 1991

003

QUADRATIC APPROXIMATIONS IN CONVEX NONDIFFERENTIABLE OPTIMIZATION*

MANLIO GAUDIOSO† AND MARIA FLAVIA MONACO†

Abstract. An implementable descent method for the unconstrained minimization of convex nonsmooth functions of several variables is described. The algorithm is characterized by the use of a set of quadratic approximations of the objective function in order to compute the search direction. The resulting direction finding subproblem is shown to be equivalent to a structured parametric quadratic programming problem. The convergence of the algorithm to the minimum is proved, and numerical experience is reported.

Key words. nondifferentiable optimization, bundle methods

AMS(MOS) subject classifications. 90C25, 65K05

1. Introduction. The research in the area of the numerical methods for nondifferentiable optimization is presently stimulated both by the relevant results that are being obtained in the field of nonsmooth analysis as well as by the practical need of algorithms performing better than those currently available.

The above motivations concurrently stimulate the research for numerical methods no longer exclusively based on piecewise linear approximations of the functions to be minimized.

In particular, the results presented in [7], [24] provide an extension to the convex nondifferentiable functions of the concept of second order derivative, through the definition of second subdifferential; on the other hand, parallel to the historical development of the numerical methods for the unconstrained minimization of smooth functions, "Newton type" methods are expected to replace "gradient type" methods [17].

This paper describes an implementable descent method for the unconstrained minimization of convex (not necessarily smooth) functions of several variables which takes advantage of the information available on the curvature of the objective function along certain directions. The algorithm is related to the family of bundle methods [12], [16], [5], the substantial difference being in the direction-finding step.

In fact, generalizing the approach presented in [6], the search direction is obtained by minimizing the directional derivative on a compact set which approximates the level set of the objective function. The approximation consists of a set of quadratic constraints generated on the basis of the "bundle" of information available.

The paper is organized as follows. In § 2 the approach is described with emphasis on the definition of the auxiliary problem to be solved at each step of the algorithm in order to find a search direction. Section 3 is devoted to the analysis of the auxiliary problem and to its reduction to a parametric quadratic program. The overall algorithm is stated in § 4. Finally the results of numerical experiments are reported in § 5.

The basic background of the paper is convex analysis, for which the fundamental reference is Rockafellar's book [20]. General references for the numerical methods are [1], [9], and [26]; in particular the bundle methods, which are strictly related to our approach, are surveyed in [14].

Standard notations are adopted throughout the paper. The symbol $\|\cdot\|$ denotes the euclidean norm. In the sequel, given a set A of the n-dimensional euclidean space,

* Received by the editors June 13, 1988; accepted for publication (in revised form) April 11, 1990.

† Dipartimento di Sistemi, Universita della Calabria 87036 Rende (CS), Italy.

we denote the convex hull of A as conv (A) and the vector of minimum norm in A (i.e., the nearest point to the origin) as Nr (A).

2. Preliminaries. Let f be a proper convex function defined on the n-dimensional euclidean space R^n. Our aim is to devise a descent method for the unconstrained minimization of f. Since the core of our approach is the determination of a search direction (possibly a descent one), we focus on this problem.

We assume that, in some iterative procedure, at a given current point $x \in R^n$, the following information is available: a subgradient vector of f at x, $g \in \partial f(x)$; a set of previously generated points x_i, $i \in F$ such that $f(x_i) > f(x)$, and the corresponding subgradients $g_i \in \partial f(x_i)$, $i \in F$.

We look for a search direction d^* obtained as a solution of a problem of the type

$$\min_d f'(x, d) \qquad h(d) \leqq 0,$$

where $f'(x, d)$ is the directional derivative of f at x along d and $h(d)$ is some reasonable model of $f(x+d)-f(x)$. In other words, the feasible region is an approximation of the level set of f at the point x, i.e., of the set

$$S(x) = \{d \mid f(x+d) \leqq f(x)\}.$$

The approach is motivated by a simple interpretation of Newton's direction in the convex quadratic case. In fact, if this is the case, a scalar multiple of Newton's direction is obtainable as a solution of

$$\min_{d \in S(x)} f'(x, d).$$

The approach has been utilized in [6] in connection with piecewise linear approximations of f, giving rise to a variant of the bundle type methods.

In this paper we will define the function $h(d)$ through explicit consideration of the curvature of f. To this aim some basic concepts are in order. Let α_i and α_i', $i \in F$ be defined as follows:

$$\alpha_i = f(x) - f(x_i) - g_i^T(x - x_i) \qquad i \in F$$

$$\alpha_i' = f(x_i) - f(x) - g^T(x_i - x) \qquad i \in F.$$

The value α_i represents the difference between $f(x)$ and the value at x of the linearization obtained starting from x_i, $i \in F$. Symmetrically, α_i' is the difference between the actual value of f and the linearization based on x, both evaluated at x_i. By convexity both α_i and α_i', $i \in F$ are nonnegative.

From the definitions, letting $d_i = x - x_i$, $i \in F$, it follows that

$$\alpha_i + \alpha_i' = (g - g_i)^T d_i \qquad i \in F$$

and, if in addition f is quadratic, say $f(x) = \frac{1}{2}x^TQx + b^Tx$, we have that $\alpha_i = \alpha_i' = \frac{1}{2}d_i^TQd_i$.

The meaning of α_i and α_i' is pictorially described in Fig. 1.

Letting $d = y - x$, $y \in R^n$, we define a set of convex quadratic functions $h_i(d)$, $i \in F$ in the following way:

$$h_i(d) = \tfrac{1}{2}\beta_i d^T d + (g_i + \beta_i d_i)^T d - \vartheta_i \qquad i \in F$$

where

$$\beta_i = (g - g_i)^T d_i / \|d_i\|^2 \quad \text{and} \quad \vartheta_i = \tfrac{1}{2}|\alpha_i - \alpha_i'|.$$

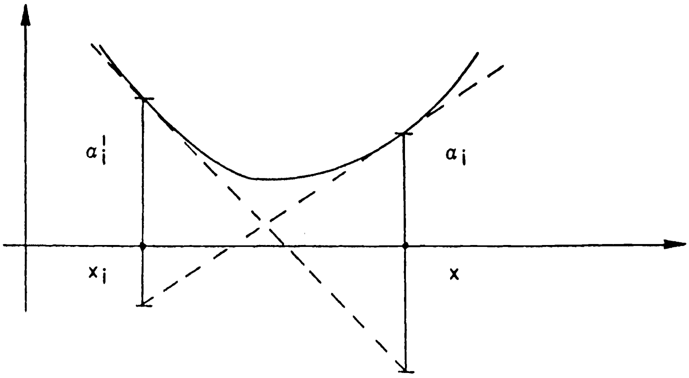


FIG. 1

The functions $h_i(d)$ are the basic elements for constructing the desired approximation of the level set. We note that the gradient of $h_i(d)$ evaluated at $d=-d_i$ is g_i and the directional derivative of $h_i(d)$, evaluated at $d=0$, along the direction d_i is equal to g^Td_i.

This means that $h_i(d)$ interpolates the first order properties of the function f at the points x_i and x, in the latter case solely along the direction d_i.

As far as the values of the functions are concerned, it is easy to verify that

$$h_i(0)=-\vartheta_i$$

$$h_i(-d_i)=\begin{cases} f(x_i)-f(x) & \text{if } \vartheta_i=\frac{1}{2}(\alpha_i-\alpha_i') \\ f(x_i)-f(x)-(\alpha_i'-\alpha_i) & \text{if } \vartheta_i=\frac{1}{2}(\alpha_i'-\alpha_i). \end{cases}$$

Note that $h_i(d)$ is a convex quadratic function whose curvature is completely defined by the value β_i obtained from information on f along d_i; its gradient at the point $d=0$, $g_i+\beta_i d_i$, can be considered as the gradient of the original function f at x_i, "transferred in a quadratic fashion" to the point x.

We define the sets S_i, $i\in F$ as

$$S_i=\{d|h_i(d)\leqq 0\}.$$

Each set S_i is consistent with the level set $S(x)$ in the sense that $S_i\cap S(x)\neq\emptyset$ (in fact at least $d=0$ is in the intersection).

We define the function $h(d)$ as follows:

$$h(d)=\max_{i\in F} h_i(d)$$

and, therefore, the approximation adopted for the level set is

$$S'=\bigcap_{i\in F} S_i=\{d|h_i(d)\leqq 0, \quad i\in F\}.$$

In conclusion, in order to find a search direction from the current point x, we solve

$$\min_d f'(x,d) \qquad h_i(d)=\tfrac{1}{2}\beta_i d^Td+(g_i+\beta_i d_i)^Td-\vartheta_i\leqq 0, \quad i\in F.$$

If the information available on the subdifferential at x is solely g, the objective function is the inner product g^Td. More in general, if a set of ε-subgradients g_i,

$g_i \in \partial_\varepsilon f(x)$, $i \in C$ of f at x are available, the directional derivative is replaced by the ε-directional derivative, i.e., the objective function becomes

$$\min_d \max_{g \in \operatorname{conv}(\Gamma_C)} g^T d$$

where Γ_C is defined as

$$\Gamma_C = \{g_i | i \in C\}.$$

Consequently the problem may be rewritten as

$$\text{(P)} \qquad \min_{d,v} v \qquad v \geqq g_i^T d \quad i \in C \qquad h_i(d) \leqq 0 \quad i \in F.$$

Before discussing the properties of (P) we just note that its feasible region is nonempty and convex.

3. The search direction finding subproblem. In this section we discuss some properties of problem (P). In the sequel we assume $\beta_i > 0$, $i \in F$ and, for simplicity of notation, we define $g_i' = g_i + \beta_i d_i$, $i \in F$.

The following proposition summarizes the properties of (P).

PROPOSITION 1. *Let S' be the set of directions d feasible for* (P) *and let* (d^*, v^*) *be the optimal solution. Then*

(a) *S' has nonempty interior unless it reduces to* $\{0\}$;

(b) $v^* \leqq 0$;

(c) *if $\vartheta_i > 0$ for all $i \in F$ then $v^* = 0$ if and only if* $0 \in \operatorname{conv}(\Gamma_C)$;

(d) *if the set $I = \{i \in F | \vartheta_i = 0\}$ is nonempty, then $v^* = 0$ is equivalent to* $0 \in \operatorname{conv}(\Gamma_C \cup \Gamma_I)$ *where* $\Gamma_I = \{g_i' | i \in I\}$.

Proof. (a) S' is the intersection of the closed balls S_i, $i \in F$

$$S_i = \{d | \, \|d + g_i'/\beta_i\|^2 \leqq r_i^2\}, \quad \text{where } r_i^2 = \|g_i'/\beta_i\|^2 + 2\vartheta_i/\beta_i.$$

S' is nonempty since each S_i contains the origin. The property follows from convexity of S' and from strict convexity of the euclidean space;

(b) consequence of feasibility of $(d, v) = (0, 0)$;

(c) the condition $\vartheta_i > 0$ for all $i \in F$ implies that S' has nonempty interior (in fact $0 \in \operatorname{int} S'$). Since $\Omega(\mathrm{P})$, the feasible region of (P), is the epigraph of the convex function $\max_{i \in C} g_i^T d$, $d \in S'$, it follows from Lemma 7.3 in [20, p. 54] that $\Omega(\mathrm{P})$ has nonempty interior, and consequently, that Slater's constraint qualification holds for (P); thus, (P) being a convex program, Kuhn-Tucker conditions are both necessary and sufficient for optimality.

The Kuhn-Tucker conditions for (P) are:

$$(1) \qquad \sum_{i \in C} \mu_i g_i + \sum_{i \in F} \pi_i g_i' + \left(\sum_{i \in F} \pi_i \beta_i \right) d = 0$$

$$(2) \qquad \mu_i \geqq 0 \qquad i \in C$$

$$(3) \qquad \pi_i \geqq 0 \qquad i \in F$$

$$(4) \qquad \sum_{i \in C} \mu_i = 1$$

$$(5) \qquad \pi_i(h_i(d)) = 0 \qquad i \in F$$

$$(6) \qquad \mu_i(v - g_i^T d) = 0 \qquad i \in C.$$

If $v^* = 0$, the direction $d = 0$ attains the optimal value. Since $0 \in \operatorname{int} S'$, from (5) the multipliers π_i must be identically zero, hence the "only if" part follows from (1).

The "if" part follows by noting that $0 \in \operatorname{conv}(\Gamma_C)$ implies that the Kuhn-Tucker conditions are satisfied by letting $d=0$, $v=0$, $\pi_i=0$, for all $i \in F$.

(d) We note first that $v^*=0$ if and only if the following system of linear inequalities has no feasible solution d:

$$\begin{aligned} &g_i^T d < 0 \qquad i \in C \\ &g_i'^T d < 0 \qquad i \in I. \end{aligned}$$

From Gordan's theorem of the alternatives [18] we have that infeasibility of the above system of linear inequalities is equivalent to feasibility of the system

$$(7) \qquad \begin{aligned} &\sum_{i \in C} \mu_i g_i + \sum_{i \in I} \pi_i g_i' = 0 \\ &\mu_i \geqq 0,\ \pi_i \geqq 0 \text{ and not all the } \mu_i \text{ and the } \pi_i \text{ equal to } 0 \end{aligned}$$

which in turn is equivalent to $0 \in \operatorname{conv}(\Gamma_C \cup \Gamma_I)$. □

The meaning of $v^*=0$ is different in (c) and (d). In particular, under the hypothesis of (c), $v^*=0$ indicates that an approximate optimality condition is verified at the current point x. Under the hypothesis of (d), $v^*=0$ may either indicate that the approximate optimality condition is verified or that the model of the level set is inadequate. More precisely, the approximation of the level set may be inherently contradictory ($S'=\{0\}$) or it may not fit with the local descent properties of f (empty intersection of S' and the set of the descent directions at x).

Our approach to the solution of (P) is based on the solution of its dual:

$$(\mathrm{D}) \qquad \begin{aligned} &\max_{\mu,\pi,d} \left(\sum_{i \in C} \mu_i g_i \right)^T d + \sum_{i \in F} \pi_i h_i(d) \\ &\sum_{i \in C} \mu_i g_i + \sum_{i \in F} \pi_i g_i' + \left(\sum_{i \in F} \pi_i \beta_i \right) d = 0 \\ &\sum_{i \in C} \mu_i = 1 \\ &\mu_i \geqq 0 \quad i \in C, \qquad \pi_i \geqq 0 \quad i \in F. \end{aligned}$$

It is easy to verify that problem (D) is always feasible. Moreover, corresponding to the feasible triplets (μ, π, d) for which $\pi \neq 0$, it is possible to express d as function of μ and π. (We note in passing that feasible triplets $(\mu, 0, d)$ exist if and only if $0 \in \operatorname{conv}(\Gamma_C)$).

Thus we eliminate d by putting

$$d = -\frac{\sum_{i \in C} \mu_i g_i + \sum_{i \in F} \pi_i g_i'}{\sum_{i \in F} \pi_i \beta_i}.$$

Consequently (D) becomes

$$\begin{aligned} &\min \frac{1}{2} \frac{\|\sum_{i \in C} \mu_i g_i + \sum_{i \in F} \pi_i g_i'\|^2}{\sum_{i \in F} \pi_i \beta_i} + \sum_{i \in F} \pi_i \vartheta_i \\ &\mu_i \geqq 0 \qquad i \in C \\ &\pi_i \geqq 0 \qquad i \in F \\ &\sum_{i \in C} \mu_i = 1. \end{aligned}$$

Defining the matrices G and G' whose columns are, respectively, the vectors g_i, $i \in C$, and g'_i, $i \in F$ and, also, the vectors μ, π, β, ϑ, $e^T = (1, 1, 1, \cdots, 1)$ of appropriate dimension, the above problem may be rewritten in compact form as follows:

$$\text{(FD)} \qquad \begin{aligned} &\min \frac{1}{2} \frac{\|G\mu + G'\pi\|^2 + \pi^T(\vartheta\beta^T + \beta\vartheta^T)\pi}{\beta^T\pi} \\ &e^T\mu = 1 \\ &\mu \geqq 0, \qquad \pi \geqq 0. \end{aligned}$$

Hence (FD) is a fractional programming problem, the numerator being a quadratic function and the denominator linear. Moreover, numerator and denominator are, respectively, convex and nonnegative on the feasible region.

In order to approach problem (FD), we define the following parametric quadratic programming problem in the nonnegative scalar parameter p

$$\text{(QP(p))} \qquad \begin{aligned} &\min \tfrac{1}{2}(\|G\mu + G'\pi\|^2 + \pi^T(\vartheta\beta^T + \beta\vartheta^T)\pi) - p\beta^T\pi \\ &e^T\mu = 1 \qquad \mu \geqq 0 \qquad \pi \geqq 0. \end{aligned}$$

PROPOSITION 2. *For positive values of the parameter p, the following statements are equivalent*:

(a) (QP(p)) *is unbounded from below*:

(b) *there exists a solution $\pi \geqq 0$, $\pi \neq 0$ to the system of linear equations*

$$G'\pi = 0, \qquad \vartheta^T\pi = 0.$$

Proof. The implication (b) implies (a) follows immediately by noting that, by hypothesis, $\beta > 0$ and that conv (Γ_C) is bounded.

To prove (a) implies (b) we note first that, because the norm of $G\mu$ is bounded on the feasible region, for all feasible μ the following holds:

$$\begin{aligned} &\tfrac{1}{2}(\|G\mu + G'\pi\|^2 + \pi^T(\vartheta\beta^T + \beta\vartheta^T)\pi) - p\beta^T\pi \\ &\qquad \geqq \tfrac{1}{2}(\|G'\pi\|^2 + \pi^T(\vartheta\beta^T + \beta\vartheta^T)\pi) + \min_{g_i \in \Gamma_C} g_i^T(G'\pi) + \tfrac{1}{2}\|u\|^2 - p\beta^T\pi \end{aligned}$$

where $u = \text{Nr}(\Gamma_C)$.

Defining $q(\pi)$ as the function on the right-hand side of the above inequality, we have that

$$q(\pi) = \min_{i \in C} q_i(\pi)$$

where, for all $i \in C$, $q_i(\pi)$ is the quadratic function

$$q_i(\pi) = \tfrac{1}{2}(\|G'\pi\|^2 + \pi^T(\vartheta\beta^T + \beta\vartheta^T)\pi) + g_i^T(G'\pi) + \tfrac{1}{2}\|u\|^2 - p\beta^T\pi.$$

From the unboundedness assumption, it follows that $q(\pi)$ must be necessarily unbounded on the feasible region. Moreover, since the cardinality of Γ_C is finite, at least one of the functions $q_i(\pi)$ must be unbounded from below. On the other hand, if (b) is not verified, all the quadratics $q_i(\pi)$ have finite minimum since the hessian matrix is strictly copositive [19]. Hence the proof follows. □

Proposition 2 indicates how to proceed to solve (FD). In fact, by solving (QP(p)) for any choice of the parameter $p > 0$, two possible outcomes may result:

(a) unbounded solution;

(b) finite minimum.

Case (a) may occur if and only if $S' = \{0\}$, which in general indicates inconsistency of the level set approximation. Whenever case (b) occur we may infer that problem

(QP(p)) has finite minimum for all the values of the nonnegative parameter p and, hence, that the function

$$\text{(QP(p))} \qquad \begin{aligned} &F(p)=\min \tfrac{1}{2}(\|G\mu+G'\pi\|^2+\pi^T(\vartheta\beta^T+\beta\vartheta^T)\pi)-p\beta^T\pi \\ &e^T\mu=1 \qquad \mu\geqq 0, \quad \pi\geqq 0 \end{aligned}$$

is well defined for $p\geqq 0$.

Further insight may be gained by determining $F(0)$. In fact different conclusions about (P) may be drawn according to $F(0)=0$ or $F(0)>0$. We observe first that $F(0)=0$ with the corresponding $\pi=0$ imply $0\in\operatorname{conv}(\Gamma_C)$; otherwise $F(0)=0$ and $\pi\neq 0$ imply the existence of a feasible solution to (7).

The implications of the case $F(0)>0$ are analyzed in the following proposition.

PROPOSITION 3. *If* $F(0)>0$ *then*:

(a) $0\notin\operatorname{conv}(\Gamma_C)$;

(b) *the system* (7) *has no solution.*

Moreover there exists a positive p^* *such that* $F(p^*)=0$ *and the corresponding* $\pi(p^*)\neq 0$.

Proof. (a) and (b) follow immediately from the definition of $F(0)$. Recalling the Kuhn-Tucker conditions for (QP(p)):

$$(8) \qquad \begin{aligned} &G^TG\mu+G^TG'\pi-\sigma e\geqq 0 \\ &G'^TG\mu+(G'^TG'+\vartheta\beta^T+\beta\vartheta^T)\pi-p\beta\geqq 0 \\ &\mu^T(G^TG\mu+G^TG'\pi-\sigma e)=0 \\ &\pi^T[G'^TG\mu+(G'^TG'+\vartheta\beta^T+\beta\vartheta^T)\pi-p\beta]=0 \\ &e^T\mu=1 \qquad \mu\geqq 0, \quad \pi\geqq 0 \end{aligned}$$

where σ is the multiplier (unconstrained in sign) of the constraint $e^T\mu=1$, we have, from (8) and from boundedness of conv (Γ_C), that, for sufficiently large values of p, the optimal pair (μ, π) cannot be of the type $(\mu, 0)$. Moreover, it is easy to show that $F(p)$ is continuous and $\lim_{p\to\infty} F(p)=-\infty$. Hence, since $F(0)>0$, we conclude that a solution p^* to the equation $F(p)=0$ exists. Finally we note that, corresponding to the optimal pair $(\mu(p^*), \pi(p^*))$, it is $\pi(p^*)\neq 0$ (otherwise the same pair should have been optimal for $p=0$, determining $F(0)=0$, which contradicts the hypothesis). □

As it will be explained in detail in the next section, the case $F(0)=0$ suggests either stopping or resetting in a possible descent procedure. On the other hand, assuming $F(0)>0$, we have the equivalence of solving the nonlinear equation $F(p)=0$ and solving the fractional programming problem (FD) (see [8], [21] for a complete treatment of the parametric approach to fractional programming).

In fact, taking into account that $\pi(p^*)\neq 0$, it is easy to show that the couple $(\mu(p^*), \pi(p^*))$ is optimal for (FD). Moreover since $F(p)$ is a concave, strictly decreasing function, the solution p^* is unique and the standard numerical methods for fractional programming apply to our case.

4. The algorithm. In this section we introduce a model descent algorithm characterized by the use of a search direction obtained through solution of problem (P). In the sequel, for the sake of simplicity of notation, we describe one iteration of the method that we define as the "main step." We assume that at x, the current estimate of the minimum, the following information is available:

— An approximation of the ε-subdifferential of f at x, for a given $\varepsilon>0$, i.e., a bundle Γ_C of subgradients of f evaluated at points "sufficiently close" to x. The bundle is assumed to contain at least g, a subgradient of f at x; in fact it reduces to the singleton after a successful line search (see step 3 below).

— The points x_i, $i \in F$ (considered "far" from x), the scalar parameters β_i, ϑ_i $i \in F$ and the related bundle of subgradients $\Gamma_F = \{g_i | i \in F\}$.

The meaning of points "close" and "far" deserves explanation. We consider close to x any point y such that, for the given $\varepsilon > 0$, each subgradient at y belongs to the ε-subdifferential of f at x. This is certainly true for all y satisfying

$$\|y - x\| \leqq t'$$

where $t' \leqq \varepsilon / 2K$ and K is any upper bound on the norm of the subgradient (see [6]). In our approach we consider "far from x" all points previously obtained in the descent procedure. Whenever $F = \emptyset$, as is the case either at the beginning of the algorithm or after a reset (see Step 1 and Step 2 below), the search direction is obtained by letting

$$d^* = -\mathrm{Nr}\,(\mathrm{conv}\,(\Gamma_C)).$$

Moreover, in this case, v^* is set to the value $-\|d^*\|^2$.

Main step. The positive parameters t', $r_{\max}$, η, m_1, m_2, $0 < m_2 < m_1 < 1$ are given.

STEP 1. Define the current parametric quadratic programming (QP(p)) and solve for $p = 0$.

Case (a) $F(0) = 0$ and $\pi^* = 0$: Stop (Optimality).

Case (b) $F(0) = 0$ and $\pi^* \neq 0$: Exit from main step (Reset F).

Case (c) $F(0) > 0$: Go to Step 2 (Finding a search direction).

STEP 2. Select a positive p and solve (QP(p)).

Case (a) Solution unbounded: Exit from main step (Reset F).

Case (b) Solve the nonlinear equation $F(p) = 0$ to obtain p^* and the corresponding optimal pair (μ^*, π^*); set

$$d^* = -\frac{G\mu^* + G'\pi^*}{\beta^T \pi^*}.$$

If $\|d^*\| > r_{\max}$, then scale d^* by putting $d^* = r_{\max}(d^* / \|d^*\|)$; set

$$v^* = -\tfrac{1}{2}\beta^T \pi^* \|d^*\|^2 - \vartheta^T \pi^*.$$

If $|v^*| < \eta$ then exit from main step (Reset F) else go to Step 3.

STEP 3. Perform a line search along d^* finding $t > 0$ and $g^+ \in \partial f(x + td^*)$ such that

(9) $$g^{+T} d^* \geqq m_1 v^*$$

and either

(10) *Case* (a) $f(x + td^*) - f(x) \leqq m_2 t v^*$

or

(11) *Case* (b) $t\|d^*\| \leqq t'$.

In Case (a) update the estimate of the minimum putting $x^+ = x + td^*$, set $\Gamma_F = \Gamma_F \cup \Gamma_C$ and calculate the new vectors β and ϑ. Moreover set $\Gamma_C = \{g^+\}$ and iterate the main step. In Case (b) set $\Gamma_C = \Gamma_C \cup \{g^+\}$ and return to Step 1.

Before stating the convergence properties of the algorithm, we briefly comment on the various steps.

We note first that, whenever Case (a) at Step 1 occurs, $0 \in \mathrm{conv}\,(\Gamma_C)$; hence the calculations are stopped because the current x is ε-optimal.

Case (b) at Step 1 indicates that no feasible descent direction exists, i.e., that possibly the local model (represented by Γ_C) and the global one (represented by Γ_F) are incongruent. Thus we cancel the information provided by the points x_i, $i \in F$, and exit from the main step. Possible satisfaction of the ε-optimality condition at the current point will be detected at the restart.

Coming to Step 2 we note that, according to Proposition 2, occurrence of Case (a) implies that Γ_F does not provide useful information in order to determine a search direction and that, on the other hand, the ε-optimality condition $0 \in \operatorname{conv}(\Gamma_C)$ is not satisfied. Hence we reset F by letting $F = \varnothing$. The scaling of d^* is to prevent the search direction from becoming arbitrarily large in norm: it has the same effect of introducing a safeguarding ball constraint in problem (P). On the other hand, the reset based on the value of $|v^*|$ is aimed to avoid shrinking phenomena of the feasible region of (P) that may occur even at points far from the minimum.

As for the line search (Step 3), the algorithm does not differ significantly from the usual line searches procedures in bundle methods [26]. Here we remark that success or failure of the line search (Case (a) and Case (b), respectively), apart from the actual modification of the current estimate of the minimum, results in different strategies for updating the bundle of information. In particular in Case (a), since a "serious step" is performed, all the information related to x is kept and the set F is updated, whereas, in Case (b) (null step), the ε-subdifferential at x is enriched by the gradient g^+.

The convergence of the algorithm is based on the following two propositions.

PROPOSITION 4. *If at any point x the algorithm generates a sequence $\{d_k^*\}$ of nondescent directions, then the corresponding sequence $\{v_k^*\}$ converges to zero.*

Proof. Suppose that a sequence of search directions has been generated and no significant descent steps have been obtained along any of them (Case (b) of the line search). Since at most only one reset may occur, it is sufficient to consider only the sequences $\{v_k^*\}$, $\{d_k^*\}$ generated after the reset. As a consequence of (9) and being $m_1 < 1$, the sequence $\{v_k^*\}$ is monotonically increasing. Moreover it is bounded from above by zero and hence it is convergent, say to v'. To prove that $v' = 0$, we note first that $\{\|d_k^*\|\}$ is bounded. Thus consider two convergent subsequences $\{d_h^*\}$, $\{v_h^*\}$, $h \in H$ and let s be the successor of h in the subsequences. Assuming that g^+ is the subgradient evaluated along d_h^*, the following hold:

$$g^{+T} d_h^* \geqq m_1 v_h^*, \qquad g^{+T} d_s^* \leqq v_s^*,$$

hence

$$g^{+T}(d_s^* - d_h^*) \leqq v_s^* - m_1 v_h^*.$$

Passing to the limit, we have $(1-m_1)v' \geqq 0$ and, since $v' \leqq 0$, we conclude that $v' = 0$, □

Proposition 4 ensures that the procedure may not remain blocked for infinitely many iterations at a point not satisfying the ε-optimality criterion. On the other hand, the following proposition guarantees that after a finite number of successful line searches the ε-optimality condition is satisfied if the objective function is bounded from below.

PROPOSITION 5. *If the function f is bounded from below and has finite minimum, for any choice of the starting point x_0 the algorithm terminates in a finite number of steps at a point satisfying the ε-optimality criterion.*

Proof. Assume that the algorithm generates an infinite subsequence $\{d_k^*\}$, $k \in K$ of directions corresponding to successful line searches. Thus we may write

$$f(x_k + t_k d_k^*) - f(x_k) \leqq m_2 t_k v_k^* \qquad k \in K$$

with

$$t_k \| d_k^* \| > t'.$$

Since f is bounded from below, we have that $\{t_k v_k^*\}$, $k \in K$ must converge to zero, which contradicts $|v_k^*| > \eta$ and $t_k > t'/r_{\max}$. Hence after a finite number of descent steps the algorithm cannot generate further descent directions and, as a consequence of Proposition 4, the current point satisfies the ε-optimality condition. □

5. Numerical experiments. We report briefly on some numerical experiments performed coding our algorithm in FORTRAN 77 (double precision) on an IBM PC.

Before discussing the results we wish to emphasize that the algorithm previously described may in fact be considered as a model algorithm, for a number of decisions may be made differently at various steps, giving rise possibly to quite distinct implementations of the basic idea. In particular, the proximity threshold t' has to be properly decided; it is necessary to specify the resetting strategy (partial resetting could be taken into consideration as well), to define appropriate maximum sizes for both Γ_C and Γ_F, to set the tolerance parameters for $F(0) = 0$ and $F(p) = 0$ etc.

The equation $F(p) = 0$ has been solved via Newton's method and the quadratic programming problem solved by the Harwell Subroutine Library code VE02AD.

The line search has been implemented by adopting a slightly modified version of Lemarechal's subroutine MLIS4.

We have considered the following standard test functions; their complete definitions may be found in the quoted references.

1. MAXQUAD [15].

$$f(x) = \max_{1 \leqq i \leqq 5} \{x^T A_i x - b_i^T x\}$$

$$n = 10, \quad f(x^*) = -0.841408, \quad x_0^T = (1, 1, \cdots, 1), \quad f(x_0) = 5337.$$

2. SHOR [25].

$$f(x) = \max_{1 \leqq i \leqq 10} \left\{ b_i \sum_{j=1}^{5} (x_j - a_{ij})^2 \right\}.$$

$$n = 5, \quad f(x^*) = 22.60016, \quad x_0^T = (0, 0, 0, 0, 1), \quad f(x_0) = 80.$$

3. ROSEN-SUZUKI [2].

$$f(x) = \max_{1 \leqq i \leqq 4} \{f_i(x)\}, \quad f_i(x) \text{ quadratic}$$

$$n = 4, \quad f(x^*) = -44, \quad x_0^T = (0, 0, 0, 0), \quad f(x_0) = 0.$$

4. CHARALAMBOUS-CONN [2].

$$f(x) = \max \{x_1^2 + x_2^4, (2 - x_1)^2 + (2 - x_2)^2, 2e^{-x_1 + x_2}\}$$

$$n = 2, \quad f(x^*) = 1.95222, \quad x_0^T = (1.0, -0.1), \quad f(x_0) = 5.41.$$

5. CHARALAMBOUS-CONN [2].

$$f(x) = \max \{x_1^4 + x_2^2, (2 - x_1)^2 + (2 - x_2)^2, 2e^{-x_1 + x_2}\}$$

$$n = 2, \quad f(x^*) = 2.0, \quad x_0^T = (2, 2), \quad f(x_0) = 20.$$

6. DEMYANOV-MALOZEMOV [3].

$$f(x) = \max\{5x_1 + x_2, -5x_1 + x_2, x_1^2 + x_2^2 + 4x_2\}$$

$$n = 2, \quad f(x^*) = -3, \quad x_0^T = (1, 1), \quad f(x_0) = 6.$$

7. MAXQ [22].

$$f(x) = \max_{1 \leqq i \leqq 20} x_i^2$$

$$n = 20, \quad f(x^*) = 0, \quad x_0^T = (1, 2, \cdots, 10, -11, -12, \cdots, -20), \quad f(x_0) = 400.$$

8. EQUILIBRIUM [15]. It is a set of three problems ($n = 5$, 8, 10) of the form:

$$f(x) = \max_{1 \leqq i \leqq n} f_i(x) \quad \text{subject to } \sum_{i=1}^{n} x_i = 1 \quad x_i \geqq 0, \qquad i = 1, \cdots, n$$

$$f(x^*) = 0, \qquad x_0^T = (1/n, 1/n, \cdots, 1/n)$$

$$n = 5, \quad f(x_0) = 20.22; \qquad n = 8, \quad f(x_0) = 9.78; \qquad n = 10, \quad f(x_0) = 26.97.$$

The above test problems are widely adopted in the literature and numerical results may be found, for example, in [2], [4], [9], [10], [11], [13], [22], [23], [25], [27].

In our experiments we have noted some degree of insensitivity to the choice of the tolerance parameter in the solution of $F(p) = 0$, (which has been set to 10^{-8}), whereas the tolerance parameter on the condition $F(0) = 0$ seems to have a stronger impact on the performance of the algorithm; in fact it not only dictates the accuracy of the solution but also influences the frequency of the resets (Cases (a) and (b) at Step 1 of the algorithm). We have adopted values ranging between 10^{-4} and 10^{-6}.

Since the test problems are all of finite minimax type, we have set the maximum size of the bundle Γ_C to the number of functions to be maximized. We have noted that a too large size of Γ_F increases the computational overhead with no apparent improvement in the overall performance (in general very few of the quadratic constraints are active at the solution of the auxiliary problem). Thus we have set the maximum size of Γ_F equal to that of Γ_C.

Our results are summarized in Table 1 where we report, for each test problem, the total number of iterations (both descent and nondescent ones), the number of function and gradient evaluations, and the function value at the stop.

On the limited number of problems tackled, the results seem comparable with those available in the literature. They demonstrate that the use of nonuniform ball constraints in the auxiliary problem is viable. On the other hand, substantial improvement in the performance would be expected by devising more sophisticated answers

TABLE 1

Test problem	Iterations	f/g eval.	$f(x_k)$	$f(x^*)$
1	40	65	−0.84140	−0.841408
2	29	59	22.60016	22.600162
3	24	53	−43.99999	−44
4	11	21	1.95222	1.952224
5	8	19	2.00000	2
6	20	41	−3.00000	−3
7	86	135	0.00000	0
8 ($n = 5$)	23	79	0.00001	0
8 ($n = 8$)	36	106	0.00006	0
8 ($n = 10$)	38	120	0.00040	0

to a number of open questions (e.g., more judicious management of the exchange of information between the bundles Γ_C and Γ_F).

6. Conclusions. We have described a model algorithm for minimizing a convex function based on the construction of suitable quadratic constraints. It may be interpreted as a bundle method embedding an automatic technique for bounding the norm of the stepsize.

The numerical experience on some standard test problems, which shows feasibility of the basic ideas underlying the approach, suggests that further research efforts should be made in order to make the proposal more competitive.

Acknowledgment. We are indebted to K. Kiwiel for suggesting some corrections to a preliminary version of the paper.

REFERENCES

[1] A. AUSLENDER, *Numerical methods for nondifferentiable optimization*, in Nonlinear Analysis and Optimization, B. Cornet, V. H. Nguyen, and J. P. Vial, eds., Math. Programming Study, 30 (1987), pp. 102–126, North-Holland, Amsterdam.

[2] J. CHARALAMBOUS AND A. R. CONN, *An efficient method to solve the minimax problem directly*, SIAM J. Numer. Anal., 15 (1978), pp. 162–187.

[3] V. F. DEMYANOV AND V. N. MALOZEMOV, *Introduction to Minimax*, John Wiley, New York, Toronto, 1974.

[4] M. FUKUSHIMA, *A descent algorithm for nonsmooth convex programming*, Math. Programming, 30 (1984), pp. 163–175.

[5] M. GAUDIOSO AND M. F. MONACO, *A bundle type approach to the unconstrained minimization of convex nonsmooth functions*, Math. Programming, 23 (1982), pp. 216–226.

[6] M. GAUDIOSO, *An algorithm for convex* NDO *based on properties of the contour lines of convex quadratic functions*, in Nondifferentiable Optimization: Motivations and Applications, V. F. Demyanov and D. Pallaschke, eds., Lecture notes in Economics and Mathematics 255, Springer-Verlag, Berlin, New York, 1985, pp. 190–196.

[7] J.-B. HIRIART-URRUTY, *A new set-valued second order derivative for convex functions*, in Fermat Days 85: Mathematics for Optimization, J.-B. Hiriart-Urruty, ed., North-Holland Mathematics Studies 129, North-Holland, Amsterdam, 1986, pp. 157–182.

[8] T. IBARAKI, *Parametric approaches to fractional programming*, Math. Programming, 26 (1983), pp. 345–362.

[9] K. C. KIWIEL, *Methods of descent for nondifferentiable optimization*, Lecture notes in Mathematics 1133, Springer-Verlag, Berlin, New York, 1985.

[10] ———, *An ellipsoid trust region bundle method for nonsmooth convex minimization*, SIAM J. Control Optim., 27 (1989), pp. 737–757.

[11] ———, *Proximity control in bundle methods for convex nondifferentiable minimization*, Math. Programming, 46 (1990), pp. 105–122.

[12] C. LEMARECHAL, *Bundle methods in nonsmooth optimization*, in Nonsmooth Optimization, C. Lemarechal and R. Mifflin, eds., Pergamon Press, Elmsford, New York, 1978, pp. 79–102.

[13] ———, *Numerical experiments in nonsmooth optimization*, in Progress in Nondifferentiable Optimization, E. A. Nurminski, ed., Report CP-82-S8, International Institute for Applied Systems Analysis, Laxenburg, Austria, 1981, pp. 61–84.

[14] ———, *Constructing bundle methods for convex optimization*, in Fermat Days 85: Mathematics for Optimization, J.-B. Hiriart-Urruty, ed., North-Holland Mathematics Studies 129, North-Holland, Amsterdam, 1986, pp. 201–240.

[15] C. LEMARECHAL AND R. MIFFLIN, *A set of nonsmooth optimization test problems*, in Nonsmooth Optimization, C. Lemarechal and R. Mifflin, eds., Pergamon Press, Elmsford, New York, 1978, pp. 151–165.

[16] C. LEMARECHAL, J. J. STRODIOT, AND A. BIHAIN, *On a bundle algorithm for nonsmooth optimization*, in Nonlinear Programming 4, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1981, pp. 245–282.

[17] C. LEMARECHAL AND J. ZOWE, *Some remarks on the construction of higher order algorithms in convex optimization*, J. Appl. Math. Optim., 10 (1983), pp. 51–68.

[18] O. L. MANGASARIAN, *Nonlinear Programming*, McGraw-Hill, New York, 1969.

[19] ———, *Solution of symmetric linear complementary problems by iterative methods*, J. Optim. Theory Appl., 22 (1977), pp. 465–485.

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

[21] S. SCHAIBLE, *Fractional programming*, Zeitschrift für Operations Res., 27 (1983), pp. 39–54.

[22] H. SCHRAMM, *Eine Kombination von Bundle- und Trust-Region-Verfahren zur Lösung nichtdifferenzierbarer Optimierungsprobleme*, Doctoral Dissertation, University of Bayreuth, 1989.

[23] H. SCHRAMM AND J. ZOWE, *A version of the bundle ideal for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results*, Report des DFG-Schwerpunktprogramms "Anwendungsbezogene Optimierung und Steuerung," University of Bayreuth, 206, 1990.

[24] A. SEEGER, *Analyse de second ordre de problemes non differentiables*, Doctoral dissertation, Univ. Paul Sabatier, Tolouse, 1986.

[25] N. Z. SHOR, *Minimization methods for non-differentiable functions*, Springer-Verlag, Berlin, New York, 1985.

[26] P. WOLFE, *A method of conjugate subgradients for minimizing nondifferentiable functions*, in Nondifferentiable Optimization, M. L. Balinski and P. Wolfe, eds., Math. Programming Study, 3 (1975), pp. 145–173, North-Holland, Amsterdam.

[27] J. ZOWE, *The* BT-*algorithm for minimizing a nonsmooth functional subject to linear constraints*, in Nonsmooth Optimization and Related Topics, F. H. Clarke, V. F. Demyanov, and F. Giannessi, eds., Plenum Press, New York, 1989, pp. 459–480.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 1, pp. 71-88, January 1991

004

AN INVERSE CONVOLUTION METHOD FOR REGULAR PARABOLIC EQUATIONS*

D. S. GILLIAM†, B. A. MAIR‡, AND C. F. MARTIN§

Abstract. In this paper, the problem of determining an unknown boundary control of a parabolic distributed parameter system, evolving over finite or infinite time, from incomplete, approximate interior temperature measurements is investigated.

First, an exact solution is obtained for square integrable boundary controls associated with a general class of parabolic equations on a finite spatial interval under the assumption that the interior temperature is known for all times.

Then, it is shown that this exact solution can be used to develop a stable algorithm for the numerical solution of this problem, without the introduction of standard regularization techniques. This algorithm assumes only knowledge of a finite, discrete set of approximate temperature readings.

One advantage of this inversion process is the availability of a priori error bounds based on the measurement errors and frequency of sampling that are obtained in this paper. Another useful feature is that it encompasses boundary controls which are arbitrary linear combinations of surface temperature and flux.

Numerical results are presented.

Key words. eigenvalue, Sturm-Liouville, convolution, Green function, iterate, spline, ill-posed, inverse heat conduction

AMS(MOS) subject classifications. 45L10, 93B07, 35R30, 65N99, 45L05

1. Introduction.

There has been considerable interest in the control, observation, and stabilization of parabolic distributed parameter systems via point sensors and actuators [8], [9]. The present work is concerned with the related question of whether or not point observation uniquely determines the boundary control for a class of parabolic boundary value problems. More precisely, the problem can be stated as follows: *From incomplete data obtained by observing the system at a given point in the spatial domain over a finite time interval, uniquely reconstruct the boundary control for this time interval within a specified class.* Varying forms of this type of problem can be found in the literature under the general heading of inverse heat conduction problems [1], [4], [6], [7], [18], [20]. The problem is often recovering the temperature and/or heat flux at the boundary from data sampled in the interior of the domain. Numerous applications of physical importance can be cast in this general setting but as is well known, such problems are ill-posed.

The technique considered here for recovering unknown surface data is based on an extension of the method developed in [12] for a very special class of problems.

* Received by the editors July 25, 1988; accepted for publication (in revised form) March 17, 1990.

† Department of Mathematics, Texas Tech University, Lubbock, Texas 79409. This research was supported in part by National Science Foundation grant DMS 8905-344, Texas Advanced Research Program grant 0219-44-5195, and by a grant from the Air Force Office of Scientific Research.

‡ Department of Mathematics, University of Florida, Gainesville, Florida 32611. This research was supported in part by National Science Foundation grant DMS 8905-344 and Texas Advanced Research Program grant 0219-44-5195.

§ Department of Mathematics, Texas Tech University, Lubbock, Texas 79409. This research was supported in part by National Science Foundation grant DMS 8905-344 and National Security Agency grant MDA 904-85-H0009.

The present method is to first compute the transfer function between the observation and control and then use a product representation for the transfer function to obtain a sequence of approximates to the control using a convolution method. It is shown that the iterates $\{f_n\}$ converge in the sense of $L^2(0, \infty)$ assuming that the sampled data derives from an L^2 control. The analysis is carried out for parabolic equations in one spatial dimension for general, regular, self-adjoint Sturm–Liouville spatial operators acting in the Hilbert state space of functions square integrable in a finite interval $[a, b]$; however, this technique can be extended to higher-dimensional rectangular bodies by using results in [13].

In addition to investigating theoretical aspects of this inverse heat conduction problem, this paper shows that the exact solution obtained in Theorem 2.1 can be used to develop a numerical solution to this problem. An important consideration here is the behavior of the inversion process in the presence of finite, noisy data. Due to the ill-posedness of this problem, the iterates obtained by using approximate data, instead of the true temperature readings, may not converge to the unknown boundary control.

Despite this unpredictable behavior of these approximations for large iterations, this paper shows how to obtain a suitable iteration level depending on the noise measurement error and frequency of measurements. Error bounds are also given for this iteration level. This problem of choosing an appropriate iteration level is similar to optimal selection problems that occur in standard regularization techniques (cf. [7], [17], [18], [20], [22]).

In § 2 we describe the basic model and state the main result. Section 3 contains a discussion of the solution of the direct problem and a useful representation for the solution of this problem as a convolution integral with kernel $K(x, t)$. In § 4 we obtain an infinite product representation for the Fourier transform of $K(x, t)$ (i.e., the system transfer function). This is in part carried out by introducing an associated Sturm–Liouville system whose spectrum corresponds to the transmission zeros of the transfer function as described in [5]. In what follows, considerable use is made of a knowledge of the properties of these transmission zeros and their relation to the eigenvalues of the original system. Also in § 4, we prove the main result stated in § 2 under the restriction that the eigenvalues and transmission zeros lie in the left half-plane, and in § 5 we remove these restrictions using a state feedback argument to shift the spectrum into the left half-plane.

We now briefly describe the inversion scheme developed in § 6. Assume that finitely many temperature readings $\{v_k\}$ are taken at the interior spatial point x_0 and at equispaced time points $\{t_k\}$ in the finite interval $[0, T]$, with error bounded above by ε and time step size h. Let v be an approximation to the true temperature u at x_0 formed by interpolating the points $\{(t_k, v_k)\}$ (e.g., by a polynomial spline). Now, by inserting v for u in the exact solution $\{f_n\}$ obtained in Theorem 2.1, we obtain a sequence of iterates $\{g_n\}$. As mentioned above, $\{g_n\}$ may be badly behaved for large n. However, we show that there is an integer $N(h, \varepsilon)$ such that $\{g_{N(h,\varepsilon)}\}$ converges to the true boundary control in $L^2[0, T]$ as h and $\varepsilon \to 0+$. In addition, a method of obtaining the value of $N(h, \varepsilon)$ is given, with the corresponding error estimates.

Finally, § 7 contains results of numerical experiments that demonstrate the appropriateness of the choice of $N(h, \varepsilon)$ and the numerical feasibility of this inversion process.

2. Notation and statement of the main result. In what follows we consider spatial and boundary operators L, B_1, B_2 given by:

$$(Lv)(x)=\frac{1}{\omega(x)}[(p(x)v'(x))'-q(x)v(x)], \qquad a<x<b$$
$$B_1(v)=h_1v(a)-k_1v'(a) \tag{2.1}$$
$$B_2(v)=h_2v(b)+k_2v'(b)$$

where prime denotes differentiation with respect to x, $p\in C^1[a,b]$; q, $\omega\in C[a,b]$, p and ω are strictly positive on $[a,b]$; h_1, k_1, h_2, and k_2 are nonnegative constants such that $h_1+k_1\neq 0$ and $h_2+k_2\neq 0$.

The present paper is concerned with the problem of determining $f(t)$, assuming:

$$\begin{aligned} &Lu=u_t, \qquad a<x<b, \qquad t>0\\ &B_1u(\cdot,t)=f(t), \qquad t>0\\ &B_2u(\cdot,t)=0, \qquad t>0\\ &u(x,0)=0, \qquad a<x<b, \end{aligned} \tag{2.2}$$

and we are able to observe the temperature values

$$\{u(x_0,t): t>0\}$$

for a fixed $x_0\in(a,b)$. Furthermore, we assume that $f\in L^2(0,\infty)$. In general, additional conditions on f are necessary to guarantee the existence of a unique solution to the direct problem of finding $u(x,t)$ given $f(t)$ in (2.2). However, here we are interested in the inverse problem of determining f assuming that $u(x,t)$ exists and that we know some of its values. One very important feature of the present method is that it can be used to recover data which is only square-integrable on finite subintervals. The reason for this is that the method only requires knowledge of the solution over a time interval $(0,T)$ to recover the unknown data for the interval $(0,T)$. Therefore, even though the results below are stated for square-integrable boundary data on $(0,\infty)$, much more general data can be considered on any finite subinterval.

Throughout this paper a central role is played by the following Sturm-Liouville systems:

$$(L-\mu_n)\psi_n=0, \quad x\in[x_0,b], \qquad \psi_n(x_0)=0, \qquad B_2\psi_n=0, \tag{2.3}$$

$$(L-\lambda_n)\varphi_n=0, \quad x\in[a,b], \qquad B_1\varphi_n=0, \qquad B_2\varphi_n=0, \tag{2.4}$$

and the function $w:[a,b]\times\mathbb{C}\to\mathbb{C}$ which satisfies

$$\begin{aligned} &(L-\lambda)w(x,\lambda)=0, \qquad a<x<b\\ &w(b,\lambda)=k_2, \qquad w'(b,\lambda)=-h_2. \end{aligned} \tag{2.5}$$

For notational convenience, we define

$$\Delta=\frac{b-a}{b-x_0}. \tag{2.6}$$

Then from standard estimates for eigenvalues of regular Sturm-Liouville systems (cf. [3]),

$$\lim_{n\to\infty}\frac{\mu_n}{\lambda_n}=\Delta^2>1. \tag{2.7}$$

Hence, there exist $\alpha\geqq 1$ and integer $m\geqq 1$ such that

$$1<\frac{\mu_n}{\lambda_n}\leqq\alpha\Delta^2, \quad \text{for all } n\geqq m. \tag{2.8}$$

Throughout this paper, α and m are considered fixed constants.

In the particular case when the spatial operator L is of the form $k(d^2/dx^2)$ for some constant k, and $k_1=k_2=0$, then $(\mu_n/\lambda_n)=\Delta^2$ for all n, so we can take $\alpha=m=1$.

With the notation from (2.1)-(2.8), we can state the main result of the paper.

THEOREM 2.1. *Assume u and f satisfy system* (2.2), $x_0\in(a,b)$ *is chosen so that* $w(x_0,0)\neq 0$, *and the values* $\{u(x_0,t)\colon t>0\}$ *are known. Also, assume that the eigenvalues* $\{\mu_n\}_{n=1}^{\infty}$, $\{\lambda_n\}_{n=1}^{\infty}$ *of the systems are strictly negative. Then the sequence f_n given by*

$$f_0(t)=\frac{h_1w(a,0)-k_1w'(a,0)}{w(x_0,0)}u(x_0,t)$$

$$f_n(t)=\frac{\mu_n}{\lambda_n}f_{n-1}(t)-\mu_n\left(1-\frac{\mu_n}{\lambda_n}\right)\int_0^t \exp\left(\mu_n(t-s)\right)f_{n-1}(s)\,ds$$

converges to f in $L^2(0,\infty)$. If in addition $f'\in L^2(0,\infty)$, then $\|f_n-f\|_2\leqq(C/n)\|f'\|_2$, for all $n\geqq m$, where C is a constant (*independent of n, f*).

In § 5, we use the result of Theorem 2.1 combined with state feedback to analyze the more general case of nonnegative eigenvalues. Also, conditions are given that ensure that $w(x_0,0)\neq 0$ holds for all $x_0\in(a,b)$. In general, the set of points for which $w(x_0,0)=0$ is a finite set.

The numerical implementation of this scheme involves the calculation of certain parameters (such as eigenvalues) of the system, which might not be known exactly. Even if $u(x_0,t)$ is known exactly, the inherent ill-posedness of the problem may lead to magnification of any errors due to approximating these parameters. However, it is shown in [14] that this inversion scheme is stable under the approximation of eigenvalues.

In § 6 we show how to compensate for its instability under noisy data by the choice of an appropriate level of iteration. Furthermore, if $w(\cdot,0)$ is not known in closed form, the values $w(a,0)$, $w'(a,0)$ and $w(x_0,0)$ may be obtained either by a series solution of the initial value problem (2.5) or by the numerical solution of the two-dimensional linear system of ordinary differential equations,

$$W'=\begin{pmatrix}0 & 1/p\\ q & 0\end{pmatrix}W,\qquad W(b)=\begin{pmatrix}k_2\\ -p(b)h_2\end{pmatrix}$$

where

$$W(x)=\begin{pmatrix}w(x)\\ p(x)w'(x)\end{pmatrix}.$$

3. The direct problem. We now describe the well-known solution of the direct problem in terms of a convolution of the control f with the inverse transform of the system transfer function, denoted by $K(x,t)$. Although many of the following results appear in a variety of sources, they are briefly repeated here for completeness and notational convenience.

From the classical theory of ordinary differential equations, we can find a function $v\colon[a,b]\times\mathbb{C}\to\mathbb{C}$ such that

$$(3.1)\qquad \begin{aligned}&(L-\lambda)v(x,\lambda)=0, && a<x<b\\ &v(a,\lambda)=k_1, && v'(a,\lambda)=h_1\end{aligned}$$

and the functions

$$\lambda\mapsto v(x,\lambda),\qquad \lambda\mapsto w(x,\lambda)$$

are entire functions of λ of order $\leqq\frac{1}{2}$ (cf. [15]).

Let

$$W(v, w)(x, \lambda) = v(x, \lambda)w'(x, \lambda) - v'(x, \lambda)w(x, \lambda)$$

be the Wronskian of v, w. Then, by a result of Abel (cf. [15]), $p(x)W(v, w)(x, \lambda)$ is independent of $x \in [a, b]$; hence, we define

$$W(\lambda) = p(x)W(v, w)(x, \lambda). \tag{3.2}$$

Therefore, W is an entire function of λ of order $\leqq \frac{1}{2}$.

Now, consider the regular Sturm-Liouville eigenvalue problem in (2.4). This system has countably many distinct eigenvalues $\{\lambda_n : n = 1, 2, \cdots\}$ such that

$$\cdots \lambda_3 < \lambda_2 < \lambda_1$$

and $\lim_{n\to\infty} \lambda_n = -\infty$ with corresponding eigenfunctions $\{\varphi_n\}$ which form a complete orthonormal basis for $L^2_\omega[a, b]$.

In this section (and the succeeding one) we assume that $\lambda_1 < 0$. In this case, the Green's function for the above system,

$$g(x, y; \lambda) = \begin{cases} \dfrac{v(x, \lambda)w(y, \lambda)}{W(\lambda)}, & a \leqq x < y \\ \dfrac{v(y, \lambda)w(x, \lambda)}{W(\lambda)}, & y \leqq x \leqq b \end{cases}$$

satisfies the following conditions.

1. The map

$$(x, y) \in [a, b] \times [a, b] \to g(x, y, \lambda)$$

is continuous for each $\lambda \in \mathbb{C}$.

2. For each $y \in [a, b]$, $\lambda \in \mathbb{C}$,
$$(L - \lambda)g(\cdot, y, \lambda) = \delta(\cdot - y), \qquad B_1 g(\cdot, y, \lambda) = B_2 g(\cdot, y, \lambda) = 0.$$

3. For each x, y, the function,

$$\lambda \mapsto g(x, y, \lambda)$$

is analytic on $\mathbb{C}$ except for simple poles at each eigenvalue λ_n.

4. The function g is represented by

$$g(x, y, \lambda) = -\sum_{n=0}^{\infty} \frac{\varphi_n(x)\varphi_n(y)}{\lambda - \lambda_n},$$

if $\lambda \neq \lambda_n$.

Now, for each x, $y \in [a, b]$, define,

$$G(x, y; t) = \begin{cases} -\sum\limits_{n=0}^{\infty} \varphi_n(x)\varphi_n(y)\, e^{\lambda_n t}, & t > 0 \\ 0, & t \leqq 0. \end{cases}$$

Then it is easily seen that for each x, y, the following conditions hold.

5. The function

$$t \mapsto G(x, y, t) \in L^1(\mathbb{R}).$$

6. The Fourier transform of $G(x, y, \cdot)$ is given by

$$\hat{G}(x, y, \xi) = g(x, y; i\xi), \qquad \xi \in \mathbb{R}. \tag{3.3}$$

7. For each y,

$$(L-i\xi)\hat{G}(\cdot, y, \xi) = \delta(\cdot - y).$$

8. For each y,

$$\left(L-\frac{\partial}{\partial t}\right)G(\cdot, y, t) = \delta(\cdot - y)\delta(t)$$

$$B_1 G(\cdot, y, t) = 0 = B_2 G(\cdot, y, t).$$

9. $G(x, y, t)$ is the fundamental solution for the parabolic equation

$$Lu = \frac{\partial u}{\partial t} \quad \text{on } [a, b] \times (0, \infty)$$

(cf. [11], [15]) satisfying

$$|G(x, y, t)| \leqq Ct^{-1/2} \exp\left(-\frac{\alpha(x-y)^2}{t}\right), \qquad t > 0 \tag{3.4}$$

and

$$\left|\frac{\partial G}{\partial x}(x, y, t)\right| \leqq Ct^{-3/2} \exp\left(-\alpha\frac{(x-y)^2}{t}\right), \qquad t > 0 \tag{3.5}$$

for strictly positive constants C, α.

From [22], the solution of the direct problem defined by the system (2.2) is given by:

$$u(x, t) = \frac{-p(a)}{h_1}\int_0^t \frac{\partial}{\partial y} G(x, y, t-s)|_{y=a} f(s)\, ds, \quad \text{if } h_1 \neq 0$$

$$u(x, t) = \frac{-p(a)}{k_1}\int_0^t G(x, a, t-s) f(s)\, ds, \quad \text{if } k_1 \neq 0.$$

Hence, if we define, for each $x \in [a, b]$, $t > 0$,

$$K(x, t) = \begin{cases} \dfrac{-p(a)}{h_1}\dfrac{\partial}{\partial y} G(x, y, t)|_{y=a}, & \text{if } h_1 \neq 0 \\ \dfrac{-p(a)}{k_1} G(x, a, t), & \text{if } k_1 \neq 0, \end{cases}$$

we have the following result.

THEOREM 3.1. *If $f \in L^2(0, \infty)$, the function $u(x, t) = K(x, \cdot) * f(t)$, is a solution of the direct problem defined by* (2.2) *and this u is the unique solution of* (2.2) *if f is continuous on* $(0, \infty)$ *with limit* 0 *at* $t = 0$.

4. Proof of the main result. In this section, we use the above representation for the solution of the direct problem to solve the inverse problem of determining f in (2.2).

Before turning to the proof of the main result, it is necessary to obtain an infinite product expansion for the Fourier transform of the function $t \to K(x_0, t)$ for any $x_0 \in (a, b)$ for which $w(x_0, 0) \neq 0$ in terms of the eigenvalues for two regular Sturm-Liouville problems.

Throughout this section, x_0 denotes a fixed point in (a, b), and $\lambda_1 < 0$.

THEOREM 4.1. *The Fourier transform,*

$$\hat{K}(x_0, \xi) = -p(a)\frac{w(x_0, i\xi)}{W(i\xi)}$$

$$= -w(x_0, i\xi)/W(v, w)(a, i\xi), \qquad \xi \in \mathbb{R}.$$

Proof. First, consider $h_1 \neq 0$. For all $t > 0$,

$$K(x_0, t) = \frac{-p(a)}{h_1} \frac{\partial}{\partial y} G(x_0, y, t)|_{y=a}.$$

By (3.5),

$$|K(x_0, t)| \leqq Ct^{-3/2} \exp\left(\frac{-\alpha(x_0 - a)^2}{t}\right), \qquad t > 0.$$

Hence, $t \mapsto K(x_0, t)$ is in $L^p(0, \infty)$, for all $p \geqq 1$. For any $\xi \in \mathbb{R}$ and $a \leqq y \leqq a + \varepsilon < x_0$, where $\varepsilon > 0$,

$$\begin{aligned}\left|\frac{\partial G}{\partial y}(x_0, y, t)\, e^{-i\xi t}\right| &\leqq \frac{C}{t^{3/2}} \frac{t^{1/4}}{(x_0 - y)^{1/2}} = Ct^{-5/4}(x_0 - y)^{-1/2} \\ &\leqq Ct^{-5/4}(x_0 - a - \varepsilon)^{-1/2}\end{aligned}$$

and

$$\int_1^\infty t^{-5/4}\, dt < \infty.$$

Hence,

$$\int_1^\infty \frac{\partial G}{\partial y}(x_0, y, t)\, e^{-i\xi t}\, dt$$

is uniformly convergent for $y \in [a, a + \varepsilon]$. This implies that,

$$\int_1^\infty K(x_0, t)\, e^{-i\xi t}\, dt = \frac{-p(a)}{h_1}\left(\frac{\partial}{\partial y} \int_1^\infty G(x_0, y, t)\, e^{-i\xi t}\, dt\right)\bigg|_{y=a}.$$

Now,

$$\int_0^1 K(x_0, t)\, e^{-i\xi t}\, dt = \frac{-p(a)}{h_1}\left(\frac{\partial}{\partial y} \int_0^1 G(x_0, y, t)\, e^{-i\xi t}\, dt\right)\bigg|_{y=a}.$$

That is,

$$\begin{aligned}\hat{K}(x_0, \xi) &= \frac{-p(a)}{h_1} \frac{\partial}{\partial y} \hat{G}(x_0, y, \xi)|_{y=a} \\ &= \frac{-p(a)}{h_1} \frac{\partial}{\partial y} g(x_0, y, i\xi)|_{y=a}, \quad \text{by (3.3)} \\ &= \frac{-p(a)}{h_1} \frac{v'(a, i\xi) w(x_0, i\xi)}{W(i\xi)} \\ &= -p(a) \frac{w(x_0, i\xi)}{W(i\xi)}, \quad \text{by (3.1)}.\end{aligned}$$

For $k_1 \neq 0$, $K(x_0, t) = -p(a)G(x_0, a, t)/k_1$. By (3.4), $t \mapsto K(x_0, t)$ is in $L^p(0, \infty)$, for all $p \geqq 1$. Hence

$$\begin{aligned}\hat{K}(x_0, \xi) &= \frac{-p(a)}{k_1} \hat{G}(x_0, a, \xi) \\ &= \frac{-p(a)}{k_1} g(x_0, a, i\xi), \quad \text{by (3.3)} \\ &= \frac{-p(a)}{k_1} \frac{v(a, i\xi) w(x_0, i\xi)}{W(i\xi)} \\ &= -p(a) \frac{w(x_0, i\xi)}{W(i\xi)}, \quad \text{by (3.1)}.\end{aligned}$$

THEOREM 4.2. *If* $w(x_0, 0) \neq 0$, *then*,

$$\hat{K}(x_0, \xi) = \frac{w(x_0, 0)}{h_1 w(a, 0) - k_1 w'(a, 0)} \frac{\prod_{n=1}^{\infty}(1 - (i\xi/\mu_n))}{\prod_{n=1}^{\infty}(1 - (i\xi/\lambda_n))}.$$

Proof. $w(x_0, z) = 0$ if and only if

$$(L - z)w(x, z) = 0, \qquad x \in [x_0, b]$$
$$w(x_0, z) = 0$$
$$h_2 w(b, z) + k_2 w'(b, z) = 0,$$

which holds if and only if there exists $n = 1, 2, 3, \cdots$ such that $z = \mu_n$. Notice that

$$-W(v, w)(a, z) = -v(a, z)w'(a, z) + v'(a, z)w(a, z)$$
$$= h_1 w(a, z) - k_1 w'(a, z).$$

Thus, $-W(v, w)(a, z) = 0$ if and only if

$$(L - z)w(x, z) = 0, \qquad x \in [a, b]$$
$$h_1 w(a, z) - k_1 w'(a, z) = 0$$
$$h_2 w(b, z) + k_2 w'(b, z) = 0$$

which is true if and only if there exists $n = 1, 2, \cdots$ such that $z = \lambda_n$.

The result follows from Hille [15], since

$$z \mapsto w(x_0, z)$$

and

$$z \mapsto W(v, w)(a, z)$$

are entire functions of order $\leqq \frac{1}{2}$, $w(x_0, 0) \neq 0$ and $\lambda_n < 0$ for all n. In fact, for all $z \in \mathbb{C}$,

$$w(x_0, z) = w(x_0, 0) \prod_{n=1}^{\infty} \left(1 - \frac{z}{\mu_n}\right)$$

$$h_1 w(a, z) - k_1 w'(a, z) = (h_1 w(a, 0) - k_1 w'(a, 0)) \prod_{n=1}^{\infty} \left(1 - \frac{z}{\lambda_n}\right).$$

Proof of Theorem 2.1 By Theorems 3.1 and 4.2,

$$\hat{f}(\xi) = \frac{h_1 w(a, 0) - k_1 w'(a, 0)}{w(x_0, 0)} \frac{\prod_{n=1}^{\infty}(1 + ia_n\xi)}{\prod_{n=1}^{\infty}(1 + ib_n\xi)} \hat{u}(x_0, \xi),$$

where $a_n = -1/\lambda_n > 0$ and $b_n = -1/\mu_n > 0$. By (2.8), $a_n > b_n$ for all $n \geqq m$. Standard estimates (cf. [3]) give $a_n = 0(n^2)$, hence $\sum_1^{\infty} a_n < \infty$ and $\sum_{n=N+1}^{\infty} a_n \leqq C/N$ for all $N \geqq m$ where C is a constant.

The result follows from an easy modification of Theorem 3.1 in [12].

5. The general case. In this section, it is shown that it can be assumed that the eigenvalues of the operator L on the interval $[a, b]$, with respect to the classical homogeneous radiation boundary conditions (2.1), all lie in the left half-plane. As has been the case thus far, we assume that the system parameters are all known so that, in particular, all the eigenvalues are known. If finitely many of these happen to be nonnegative, we show that by using state feedback, it is possible to reconstruct the original unknown boundary input f by a modification of our main result.

If the eigenvalues for the systems (2.3), (2.4) are all strictly negative, define $\eta = 0$. If not, choose $\eta > 0$ such that for every μ, λ satisfying (2.3), (2.4), respectively, we have

$$\mu < \eta \quad \text{and} \quad \lambda < \eta.$$

Then the systems:

$$(5.1)\qquad \begin{aligned} &Lu = u_t, \qquad a<x<b, \qquad t>0\\ &B_1 u(\cdot, t) = f(t), \qquad t>0\\ &B_2 u(\cdot, t) = 0, \qquad t>0\\ &u(x,0) = 0, \qquad a<x<b \end{aligned}$$

and

$$(5.2)\qquad \begin{aligned} &(L-\eta)U = U_t, \qquad a<x<b, \qquad t>0\\ &B_1 U(\cdot, t) = e^{-\eta t} f(t), \qquad t>0\\ &B_2 U(\cdot, t) = 0, \qquad t>0\\ &U(x,0) = 0, \qquad a<x<b \end{aligned}$$

are equivalent if $U(x, t) = e^{-\eta t} u(x, t)$.

Also note that if w_η satisfies

$$(5.3)\qquad \begin{aligned} &((L-\eta)-\lambda) w_\eta(x,\lambda) = 0, \qquad a<x<b\\ &w_\eta(b,\lambda) = k_2, \qquad w_\eta'(b,\lambda) = -h_2 \end{aligned}$$

then $w_\eta(x, \lambda) = w(x, \eta + \lambda)$.

Let $\{\mu_n\}_{n=1}^\infty$ and $\{\lambda_n\}_{n=1}^\infty$ be as defined earlier in (2.3), (2.4).

Now, for $\eta > \lambda_1$ and $w(x_0, \eta) \neq 0$, define for $t>0$ the sequence $f_n^\eta(t)$ by:

$$f_0^{(\eta)}(t) = \frac{h_1 w(a,\eta) - k_1 w'(a,\eta)}{w(x_0,\eta)}\, U(x_0, t)$$

$$f_n^{(\eta)}(t) = \frac{\mu_n - \eta}{\lambda_n - \eta} f_{n-1}^{(\eta)}(t) - (\mu_n - \eta)\left(1 - \frac{(\mu_n-\eta)}{(\lambda_n - \eta)}\right) \exp\left((\mu_n - \eta)t\right) * f_{n-1}^{(\eta)}(t).$$

Then, $\{f_n^{(\eta)}\}$ converges to $f^{(\eta)} = e^{-\eta t} f(t)$ in $L^2(0,\infty)$ and if $(f^{(\eta)'}) \in L^2(0, \infty)$, then

$$\|f_n^{(\eta)} - f^{(\eta)}\|_2 = O\left(\frac{1}{n}\right) \quad \text{as } n \to \infty.$$

In order to apply the inversion process it is required that

$$w(x_0, \eta) = w_\eta(x_0, 0) \neq 0.$$

It is therefore of interest to know exactly what restriction this places on the choice of x_0. The following remark indicates that under a variety of conditions, there is no restriction on the choice of x_0, or, at most one point must be omitted. In general, the set of x_0 that must be excluded is a finite set. We note the following special cases.

1. If q in (2.1) is identically zero, then $w(x_0, \eta) \neq 0$ for all $x_0 \in (a, b)$.
2. If q is strictly positive, then

$$w(x_0, \eta) \frac{\partial w}{\partial x}(x_0, \eta) = 0$$

for at most one $x_0 \in [a, b]$ (cf. [15, p. 437]). Hence, in this case if h_1 or k_1, or h_2 or k_2 is zero, then $w(x_0, \eta) \neq 0$, for all $x_0 \in (a, b)$. Also, if q is strictly postive and h_2 and k_2 are not zero (or h_1 and k_1 are not zero), then $w(x_0, \eta) = 0$ for at most one $x_0 \in (a, b)$.

6. The inversion scheme. Here we describe and analyze a technique for the numerical implementation of the inversion scheme obtained in § 4. Numerical evidence of the success of this method in dealing with both surface temperature and heat flux is presented in the next section. An interesting and very useful feature of this technique, which is obtained here, is the availability of an estimate on an appropriate iteration level which depends on the accuracy of the available temperature readings. The method described here assumes only the knowledge of a finite, discrete set of approximate values of the temperature at an interior point x_0.

Throughout this section we assume that the system (2.2) is being observed for a finite time T, so the boundary control f is defined only on $[0, T]$. We extend f to all of $\mathbb{R}$ by defining it to be zero off $[0, T]$. Let x_0 be fixed in (a, b) such that $w(x_0, 0) \neq 0$, and $\mu_1, \lambda_1 < 0$. Denote the true temperature $u(x_0, t)$ by $u(t)$.

Now assume that the temperature is observed at a discrete set of times $\{t_1, t_2, \cdots, t_M\}$, where $t_k = kh$, $h > 0$ is fixed, and $t_M = T$.

Let v_k be the approximate temperature reading at time t_k, $k = 1, 2, \cdots, M$, with measurement error given by

$$|v_k - u(t_k)| \leqq \varepsilon, \quad \text{for } k = 1, 2, \cdots, M. \tag{6.1}$$

For notational convenience, let $t_0 = 0$, $v_0 = 0$.

Due to the accumulation of round-off errors, it is more efficient to replace the iterative generation of the sequence $\{f_n\}$ of approximations to the boundary control f by the following formula:

$$f_N(t) = A\left(\prod_{n=1}^{N} \frac{\mu_n}{\lambda_n}\right)\left\{u(t) + \sum_{n=1}^{N} \gamma_{N,n} \exp(\mu_n t) * u(t)\right\} \tag{6.2}$$

where

$$A = \frac{h_1 w(a, 0) - k_1 w'(a, 0)}{w(x_0, 0)},$$

$$\gamma_{N,n} = \prod_{k=1}^{N} (\mu_n - \lambda_k) \Big/ \prod_{\substack{k=1 \\ k \neq n}}^{N} (\mu_n - \mu_k), \qquad n = 1, \cdots, N.$$

This result is easily obtained by a partial fraction decomposition of the Fourier or Laplace transform of f_N.

Now the first step in this numerical procedure is to generate an approximation, v, to the true temperature u, such that

$$v(t_k) = v_k, \qquad k = 0, 1, 2, \cdots, M. \tag{6.3}$$

In the numerical examples presented here, v was chosen to be the cubic spline with the "not-a-knot" condition (cf. [10]). However, higher degree polynomial splines may also be used. Since v_k is only approximately equal to $u(t_k)$, to analyze the error between u and v, we introduce an intermediate function σ.

LEMMA 6.1. *For each* δ, $0 < \delta < h$, *there exists* $\sigma \in \mathscr{C}^{\infty}(\mathbb{R})$ *such that*:

(a) $\sigma(t_k) = v_k$, *for* $k = 0, 1, \cdots, M$.

(b) $\sigma(t) = 0$ *for* $t \leqq 0$ *and* $\sigma(t) = u(t)$ *for* $t \geqq T + \delta/2$.

(c) $\|\sigma - u\|_2 < T^{1/2}\varepsilon$.

(d) $\|\sigma^{(n)}\|_\infty \leqq C_n(\varepsilon/\delta^n + \|f\|_2)$, *for all* $n = 1, 2, \cdots$, *where* C_n *is a constant* (*independent of* δ, h, ε).

Proof. Choose $\Psi \in \mathscr{C}^\infty(\mathbb{R})$ such that

(i) $0 \leqq \Psi(t) \leqq 1$, for all $t \in \mathbb{R}$,

(ii) $\Psi^{(n)}(t) = 0$, for all $|t| \geqq 1$ and $n = 0, 1, 2, \cdots$,

(iii) $\|\Psi\|_\infty = \Psi(0) = 1$.

For each $k = 1, 2, \cdots, M$, define,

$$J_k = (t_k - \delta/2, t_k + \delta/2),$$

and

$$\Psi_k(t) = \Psi\left(\frac{2}{\delta}(t - t_k)\right), \quad \text{for } t \in J_k.$$

Let $\sigma = \sum_{k=1}^M (v_k - u(t_k))\Psi_k + u$.

Then, by the disjointness of $\{J_k : k = 1, 2, \cdots, M\}$,

$$\sigma(t) = \begin{cases} (v_k - u(t_k))\Psi_k(t) + u(t), & \text{if } t \in J_k \\ u(t), \quad \text{if } t \notin \bigcup\limits_{k=1}^{M} J_k. \end{cases}$$

From Theorem 3.1 and [14], we have that,

$$u = K(x_0, \cdot) * f,$$

where

$$|\hat{K}(x_0, \xi)| \leqq C \exp\left\{-\left(\frac{x_0 - a}{2}\right)\sqrt{\frac{|\xi|}{2}}\right\},$$

for all $\xi \in \mathbb{R}$, for some constant C.

Hence the Sobolev embedding theorem implies $u \in \mathscr{C}^\infty(\mathbb{R})$ and Young's inequality gives

$$\|u^{(n)}\|_\infty \leqq \|K^{(n)}(x_0, \cdot)\|_2 \|f\|_2.$$

It then follows that $\sigma \in \mathscr{C}^\infty(\mathbb{R})$ and $\|u^{(n)}\|_\infty \leqq C_n \|f\|_2$.

(a) and (b) of Lemma 6.1 are easily verified.

$$\begin{aligned} \|\sigma - u\|_2^2 &= \sum_{k=1}^M \int_{J_k} (v_k - u(t_k))^2 \Psi_k^2(t)\, dt \\ &\leqq \varepsilon^2 M \frac{\delta}{2} \|\Psi\|_2^2 < \varepsilon^2 T. \end{aligned}$$

If $t \in J_k$, then

$$\begin{aligned} |\sigma^{(n)}(t)| &\leqq |v_k - u(t_k)|\,|\Psi_k^{(n)}(t)| + |u^{(n)}(t)| \\ &\leqq \varepsilon\left(\frac{2}{\delta}\right)^n |\Psi^{(n)}(t)| + C_n \|f\|_2 \\ &\leqq C_n\left(\frac{\varepsilon}{\delta^n} + \|f\|_2\right). \end{aligned}$$

If $t \notin \bigcup_{k=1}^M J_k$, $|w^{(n)}(t)| = |u^{(n)}(t)| \leqq C_n \|f\|_2$.

This completes the proof.

Consequently, v is an interpolant of the points $(t_k, \sigma(t_k))$ on the graph of the $\mathscr{C}^\infty$ function σ, hence, depending on the other imposed conditions on v, we may have an estimate of the form:

$$\|v-\sigma\|_2 \leqq C_p \|w^{(p)}\|_\infty h^p \tag{6.4}$$

for some integer p and constant $C_p > 0$.

Throughout the remainder of this paper, C denotes a constant which is not necessarily the same at each occurrence, and p is a fixed constant.

THEOREM 6.1. *If v satisfies* (6.4) *then*

$$\|v-u\|_2 \leqq C(\varepsilon + h^p \|f\|_2).$$

Proof. By choosing $n = p$ and $\delta = h/2$ in Lemma 6.1(d) and using (6.4), we obtain

$$\|v-\sigma\|_2 \leqq C(\varepsilon + h^p \|f\|_2).$$

The result follows from the triangle inequality and Lemma 6.1(c).

The strategy for the numerical implementation of Theorem 2.1 is to compute approximations to f by using the formula (6.2) where v is inserted for the unknown function u. These approximations will be denoted g_N. So,

$$g_N(t) = A\left(\prod_{n=1}^{N} \frac{\mu_n}{\lambda_n}\right)\left\{v(t) + \sum_{n=1}^{N} \gamma_{N,n} \exp(\mu_n t) * v(t)\right\}. \tag{6.5}$$

Since v is "close" to u, g_N is "close" to f_N for each N, however, due to the ill-posed nature of the problem, $\{g_N\}$ usually will not converge to f. It will now be shown how to overcome this difficulty by choosing an appropriate level of iteration.

LEMMA 6.2. *Let $r > 0$, $s > 0$. If*

$$h \leqq \exp(-(2\alpha\Delta^2 - 1)^{(m+1)/r})$$

and

$$\varepsilon \leqq \exp(-(2\alpha\Delta^2 - 1)^{(m+1)/s}),$$

then there exist positive integers $N_r(h)$, $M_s(\varepsilon)$, and $N_{r,s}(h, \varepsilon)$ such that

$$\frac{r \log\log(1/h)}{\log(2\alpha\Delta^2 - 1)} - 1 \leqq N_r(h) \leqq \frac{r \log\log(1/h)}{\log(2\alpha\Delta^2 - 1)},$$

and, if $\varepsilon > 0$,

$$\frac{s \log\log(1/\varepsilon)}{\log(2\alpha\Delta^2 - 1)} - 1 \leqq M_s(\varepsilon) \leqq \frac{s \log\log(1/\varepsilon)}{\log(2\alpha\Delta^2 - 1)},$$

$$N_{r,s}(h, \varepsilon) = \min(N_r(h), M_s(\varepsilon)),$$

and, $N_{r,s}(h, 0) = N_r(h)$, and

$$\|g_{N_{r,s}(h,\varepsilon)} - f\|_2 \leqq C\left(\frac{2\alpha\Delta^2 + 1}{2\alpha\Delta^2 - 1}\right)^{m-1} \{h^p(\log 1/h)^r \|f\|_2 + \varepsilon(\log 1/\varepsilon)^2\}$$
$$+ \|f_{N_{r,s}(h,\varepsilon)} - f\|_2.$$

Proof. By using the recurrence relations satisfied by $\{f_n\}$ and $\{g_n\}$, Young's inequality, and (2.7), it can be shown that

$$\|g_n - f_n\|_2 \leqq C\left(\frac{2\alpha\Delta^2 + 1}{2\alpha\Delta^2 - 1}\right)^{m-1} (2\alpha\Delta^2 - 1)^n \|v - u\|_2, \quad \text{for } n \geqq m,$$

(cf. [16]). Hence, by Theorem 6.1,

$$\|g_n - f\|_2 \leqq C\left(\frac{2\alpha\Delta^2+1}{2\alpha\Delta^2-1}\right)^{m-1}(2\alpha\Delta^2-1)^n(\varepsilon + h^p\|f\|_2) + \|f_n - f\|_2.$$

Now, by the assumptions on h and ε,

$$m \leqq \frac{r \log\log(1/h)}{\log(2\alpha\Delta^2-1)} - 1$$

and

$$m \leqq \frac{s \log\log(1/\varepsilon)}{\log(2\alpha\Delta^2-1)} - 1.$$

So we can choose $N_r(h)$, $M_s(\varepsilon)$ and $N_{r,s}(h, \varepsilon)$ as in the statement of the lemma, and $N(h, \varepsilon) \geqq m$. The result follows by noting that

$$(2\alpha\Delta^2-1)^{N_{r,s}(h,\varepsilon)} \leqq \min\{(\log 1/h)^r, (\log 1/\varepsilon)^s\}.$$

Now, as ε, h approach zero, there must exist r, s such that h, ε, r, s satisfy the conditions in Lemma 6.2. In the following, we assume such r, s are fixed and let h, ε tend to zero.

Let H^1 denote the Sobolev space $\{f \in L^2 : f' \in L^2\}$ and $\|\cdot\|_{H^1}$ the norm on H^1.

THEOREM 6.2. *Let h, ε and $N_{r,s}(h, \varepsilon)$ be as in Lemma 6.2. Then,*

$$\lim_{h,\varepsilon \to 0+} g_{N_{r,s}(h,\varepsilon)} = f, \quad \text{in } L^2.$$

Furthermore, for all f in any given ball, $\{f \in H^1 : \|f\|_{H^1} \leqq D\}$,

$$\begin{aligned}\|g_{N_{r,s}(h,\varepsilon)} - f\|_2 &= O(h^p(\log 1/h)^r + \varepsilon(\log 1/\varepsilon)^s) \\ &\quad + O\left(\frac{1}{\log\log 1/h} + \frac{1}{\log\log 1/\varepsilon}\right) \\ &= O\left(\frac{1}{\log\log 1/h} + \frac{1}{\log\log 1/\varepsilon}\right).\end{aligned}$$

Proof. The first assertion follows from Lemma 6.2 and from the first assertion in Theorem 2.1.

For the second assertion, note that,

$$\frac{1}{N_{r,s}(h, \varepsilon)} \leqq \frac{1}{N_r(h)} + \frac{1}{M_s(\varepsilon)} = O\left(\frac{1}{\log\log 1/h} + \frac{1}{\log\log 1/\varepsilon}\right).$$

The proof then follows from Lemma 6.2 and Theorem 2.1.

Now, the conditions on on ε, h, r, s in Lemma 6.2 only require r and s to be bounded below. Hence, for given ε and h, $N_{r,s}(h, \varepsilon)$ may be chosen arbitrarily large. However, Theorem 6.2 indicates that the error between $g_{N_{r,s}(h,\varepsilon)}$ and f may become unbounded as r, $s \to \infty$. To make an appropriate choice for r, s it seems reasonable to require that the error be minimized.

From the proof of Lemma 6.2 and Theorem 6.2, this leads to the problem of minimizing

$$H(r)=C_m(\Delta)h^p(\log 1/h)^r+\frac{\log(2\alpha\Delta^2-1)}{r\log\log 1/h},\quad \text{subject to}$$

$$r\geqq r_0=(m+1)\frac{\log(2\alpha\Delta^2-1)}{\log\log 1/h};\quad \text{and}$$

$$E(s)=C_m(\Delta)\varepsilon(\log 1/\varepsilon)^s+\frac{\log(2\alpha\Delta^2-1)}{s\log\log 1/\varepsilon},\quad \text{subject to}$$

$$s\geqq s_0=(m+1)\frac{\log(2\alpha\Delta^2-1)}{\log\log 1/\varepsilon},\quad \text{where}$$

$$C_m(\Delta)=A\left(\frac{2\alpha\Delta^2+1}{2\alpha\Delta^2-1}\right)^{m-1}.$$

By elementary calculations, the minimum of H occurs either at r_0 or at r_1 for which $H'(r_1)=0$ and $r_0\leqq r_1$. By using the equation $H'(r_1)=0$ and the inequality $r_1\geqq r_0$, it can be seen that the minimum of H occurs at a point in the interval

$$[r_0,\{h^pC_m(\Delta)(m+1)^2\log(2\alpha\Delta^2-1)\log\log 1/h\}^{-1}]. \tag{6.6}$$

A similar analysis for $E(r)$ shows that the minimum of E occurs at a point in the interval

$$[s_0,\{\varepsilon C_m(\Delta)(m+1)^2\log(2\alpha\Delta^2-1)\log\log 1/\varepsilon\}^{-1}]. \tag{6.7}$$

The determination of the exact location of these minima requires the solution of a highly nonlinear equation which cannot be solved exactly. However, by using the bounds in (6.6), (6.7) and the convexity of the functions H and E, it is readily shown that the Newton-Raphson technique (applied to the location of the zeros of H' and E') enables us to numerically determine the location of these minima to any desired degree of accuracy.

Using these approximate values of r and s in Lemma 6.2, we obtain an approximate level of iteration which guarantees the stability of the inversion process. In the numerical results presented later, these iteration levels are referred to as *recommended iterates.*

This demonstrates the feasibility of the proposed inversion process given only finite, discrete, noisy temperature readings.

We conclude this section by examining the problem of the computation of the convolutions appearing in the inversion scheme. To overcome this, it is suggested that the approximation v be chosen to be a polynomial spline of some degree q. Then the convolutions can be calculated exactly as follows.

$$v(t)=\sum_{j=1}^{q}C_{ij}(t-t_i)^j,\quad \text{for } t\in[t_i,t_{i+1}],$$

$i=0,1,2,\cdots,M-1$, where C_{ij} are known.

$$\begin{aligned}\exp(\mu_nt)*v(t)|_{t=t_k}&=\sum_{i=0}^{k-1}\int_{t_i}^{t_{i+1}}\exp(\mu_n(t_k-s))v(s)\,ds\\&=\sum_{i=0}^{k-1}\exp(\mu_nh(k-i))\sum_{j=1}^{p}C_{ij}I_{nj},\end{aligned}$$

where

$$I_{n1}=\frac{1-\exp(-\mu_n h)}{\mu_n}$$

$$I_{n(j+1)}=-\frac{h^j}{\mu_n}\exp(-\mu_n h)+\frac{j}{\mu_n}I_{nj}.$$

Finally, this inversion scheme can be extended to deal with higher-dimensional rectangular bodies. For example, consider the following system on a two-dimensional rectangle.

$$\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}=\frac{\partial u}{\partial t},\qquad 0<x<a,\qquad 0<y<b,\qquad t>0$$

$$u(x,y,0)=0,\qquad 0\leqq x\leqq a,\qquad 0\leqq y\leqq b$$

$$u(x,0,t)=u(x,b,t)=0,\qquad 0\leqq x\leqq a,\qquad t\geqq 0$$

$$u(a,y,t)=0,\qquad 0\leqq y\leqq b,\qquad t\geqq 0$$

$$u(0,y,t)=f(y,t),\qquad 0\leqq y\leqq b,\qquad t\geqq 0.$$

Then, the following integral representation theorem is well known (cf. [13]):

$$u(x,y,t)=\int_0^t \varphi_a(x,t-s)\int_0^b G_b(y,t,\xi,s)f(\xi,s)\,d\xi\,ds,$$

where φ_a and G_b are Green's functions for corresponding one-dimensional problems.

In [13], it was shown that for each fixed $x_0\in(0,a)$, the values of $u(x_0,y,t)$ for $y\in(0,b)$, $t>0$, uniquely determine the boundary function f. Due to the form of the kernel appearing in the above integral representation, Theorem 2.1 extends naturally to a technique for recovering the Fourier coefficients

$$\tilde{f}_m(t)=\int_0^b \psi_m(y)f(y,t)\,dy,$$

thus determining f by $f(y,t)=\sum\tilde{f}_m(t)\psi_m(y)$.

7. Numerical results. In this section, we present the results of numerical tests of the inversion scheme described in § 6 for surface temperature and flux, boundary conditions in the cases when the boundary function is the characteristic function of [0.25, 0.5] and $\sin(2\pi t)$. The system is observed at discrete, equispaced points over the interval $[0,T]$, $T=1$ or $T=2$. Our results demonstrate the appropriateness of the choice of the so-called *recommended iterates*, for an oscillatory boundary control which lies in the Sobolev space H^1. The effectiveness of this method is also demonstrated for the characteristic function boundary control, although our mathematical analysis does not seem to cover this case.

In all the results presented here, the following values are fixed.

$$a=0,\qquad b=\pi,\qquad x_0=0.1,\qquad h_2=1,\qquad k_2=0,$$

and

$$L\equiv\frac{d^2}{dx^2}.$$

As it is not possible to obtain exact solutions to these direct problems for the finite rod, the data was generated by using a series solution and taking enough terms to guarantee an accuracy of 10^{-4} at each sampling point

$$(x_0, t_k).$$

Then random noise was added to these approximate values with standard deviation the magnitude of *std. dev.* as indicated in each figure below. The quantity *error* indicated in each diagram is the $L^2(0, T)$ error between the indicated *iterate* and the true boundary function.

In all the figures below, the approximation produced by the inversion scheme is in solid line and the true boundary function appears as a dotted graph.

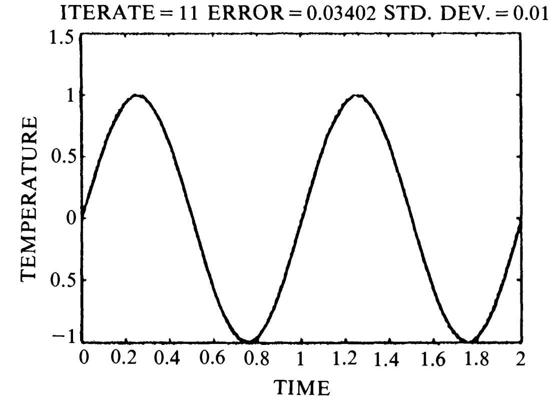
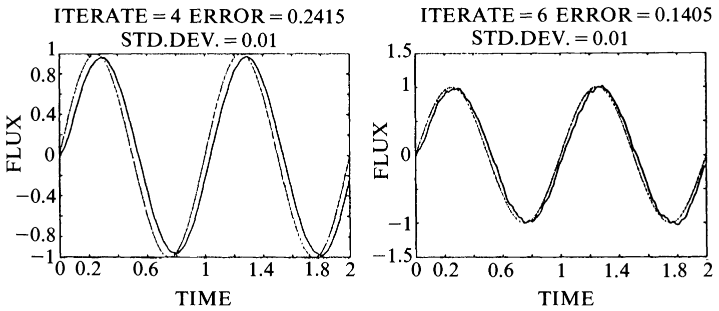
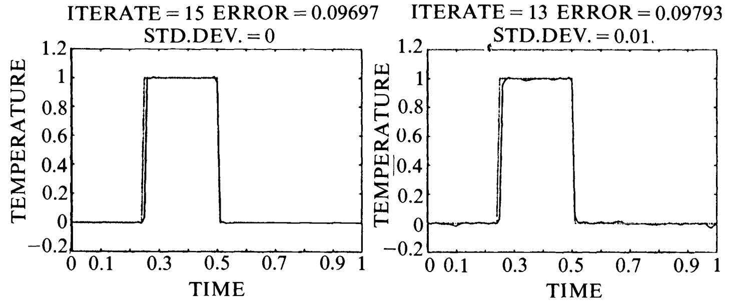

ITERATE = 11 ERROR = 0.03402 STD. DEV. = 0.01

FIG. 1. *Sine surface temperature,* 11 *iterates.*


ITERATE = 4 ERROR = 0.2415 STD.DEV. = 0.01

ITERATE = 6 ERROR = 0.1405 STD.DEV. = 0.01

FIG. 2. *Sine heat flux with time step size* 0.01.


ITERATE = 15 ERROR = 0.09697 STD.DEV. = 0

ITERATE = 13 ERROR = 0.09793 STD.DEV. = 0.01

FIG. 3. *Discontinuous surface temperature.*

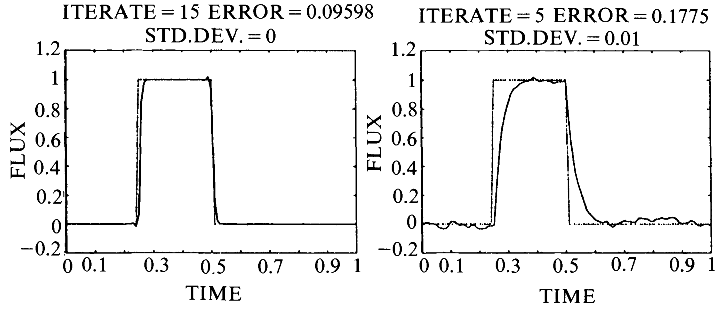


FIG. 4. *Discontinuous surface heat flux.*

Example 7.1. In this case, $h_1 = 1$, $k_1 = 0$, $f(t) = \sin(2\pi t)$ on $[0, 2]$ and time step size, $h = 0.04$. The inversion scheme produced Table 1 and Fig. 1.

We now present a result for the recovery of flux (Table 2, Fig. 2).

Example 7.2. Here, $h_1 = 0$, $k_1 = 1$, $f(t)$ as in Example 7.1 and time step size, $h = 0.01$.

The next two examples (see Figs. 3, 4) present data on the recovery of f in the case when

$$f(t) = \begin{cases} 1, & \frac{1}{4} \leqq t \leqq \frac{1}{2} \\ 0, & \text{otherwise.} \end{cases}$$

Although the analysis in the paper does not provide explicit error bounds for this case, the numerical results are surprisingly good. In both examples the time step size is 0.01.

These numerical results compare favorably with those in [7], [18]. In this regard, it is important to note that the problem was analyzed in [7], [18], for a semi-infinite rod, whereas our results are valid for a rod of finite length.

TABLE 1. *Data for Fig.* 1.

Recommended iterate = 11
Predicted error = 0.2000

Iterate	9	11	15	17	19	21
Error	0.0406	0.0340	0.0288	0.0295	0.0321	0.0365

TABLE 2. *Data for Fig.* 2.

Recommended iterate = 4
Predicted error = 1.0900

Iterate	3	4	6	8	10	12
Error	0.3482	0.2415	0.1405	0.1140	0.1420	0.2120

REFERENCES

[1] J. V. BECK, B. BLACKWELL, AND C. R. ST. CLAIR, JR., *Inverse Heat Conduction, Ill-posed Problems*, John Wiley, New York, 1985.

[2] J. V. BECK, *Nonlinear estimation applied to the nonlinear heat conduction problem*, Int. J. Heat Mass Trans., 13 (1970), pp. 703-716.

[3] G. BIRKHOFF AND G.-C. ROTA, *Ordinary Differential Equations*, Blaisdell, New York, 1969.

[4] O. R. BURGGRAF, *An exact solution of the inverse problem in heat conduction theory and applications*, ASME J. Heat Transfer, 86c (1964), pp. 373-382.

[5] C. I. BYRNES, *Adaptive stabilization of infinite dimensional systems*, in Proceedings of 26th IEEE Conference on Decision and Control, IEEE Control Systems Society, New York, 1987, pp. 1435-1440.

[6] J. R. CANNON, *The one-dimensional heat equation*, in Encyclopedia of Mathematics and its Applications, Vol. 23, Addison-Wesley, New York, 1984.

[7] A. CARASSO, *Determining surface tempeatures from interior observations*, SIAM J. Appl. Math., 42 (1982), pp. 558-574.

[8] R. F. CURTAIN AND A. J. PRITCHARD,*Infinite Dimensional Linear System Theory*, Lecture Notes in Computer and Information Science 8, Springer-Verlag, Berlin, New York, 1978.

[9] R. F. CURTAIN, *Finite dimensional compensators for parabolic distributed systems with unbounded control and observation*, SIAM J. Control Optim., 22 (1984), pp. 255-276.

[10] CARL DE BOOR, *A Practical Guide to Splines*, Applied Math. Sciences, vol. 27, Springer-Verlag, Berlin, New York, 1978.

[11] A. FRIEDMAN, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.

[12] D. S. GILLIAM, B. A. MAIR, AND C. F. MARTIN, *A convolution method for inverse heat conduction problems*, Math. Systems Theory, 21 (1988), pp. 49-60.

[13] ———, *Observability and determination of surface temperature, Part* 1, Internat. J. Control, 48 (1988), pp. 2249-2264.

[14] D. S. GILLIAM AND B. A. MAIR, *Stability of a convolution method for inverse heat conduction problems*, submitted.

[15] E. HILLE, *Lectures on Ordinary Differential Equations*, Addison-Wesley, New York, 1969.

[16] B. A. MAIR, *On the recovery of surface temperature and heat flux via convolutions*, in Computation and Control, Birkhäuser, Boston, 1989, pp. 197-208.

[17] V. A. MOROZOV, *Methods for Solving Incorrectly Posed Problems*, Springer-Verlag, Berlin, New York, 1984.

[18] D. A. MURIO, *Parameter selection by discrete mollification and the numerical solution of the inverse heat conduction problem*, J. Comp. Appl. Math., 22 (1988), pp. 25-34.

[19] D. A. MURIO AND C. C. ROTH, *An integral solution for the inverse heat conduction problem after the method of Weber*, Comput. Math. Appl. 15 (1988), pp. 39-51.

[20] A. N. TIKHONOV AND V. Y. ARSENIN, *Solutions of Ill-Posed Problems*, John Wiley, New York, 1971.

[21] J. M. VARAH, *A practical examination of some numerical methods for linear discrete ill-posed problems*, SIAM Rev., 21 (1979), pp. 100-111.

[22] E. ZAUDERER, *Partial Differential Equations of Applied Mathematics*, Wiley-Interscience, New York, 1983.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 1, pp. 89-118, January 1991

INITIAL BOUNDARY VALUE PROBLEMS AND OPTIMAL CONTROL FOR NONAUTONOMOUS PARABOLIC SYSTEMS*

P. ACQUISTAPACE,† F. FLANDOLI,‡ AND B. TERRENI§

Abstract. A large class of linear nonautonomous parabolic systems in bounded domains is considered, with control acting on the boundary through Dirichlet or Neumann conditions, from the point of view of semigroup theory. The results from [*Rend. Sem. Mat. Univ. Padova*, 78 (1987), pp. 47-107], [*On fundamental solutions for abstract parabolic equations*, Lecture Notes in Math., Vol. 1223, Springer-Verlag, Berlin, Heidelberg, 1986, pp. 1-11] on abstract homogeneous parabolic Cauchy problems allow operators with varying domains and Hölder continuous coefficients to be handled. A representation formula for solutions corresponding to square integrable control functions is derived and used to solve a linear-quadratic regulator problem over finite time horizon, by a direct study of the associated integral Riccati equation.

Key words. optimal control, parabolic systems, boundary control, Riccati equation

AMS(MOS) subject classifications. 49A22, 49B22, 34H05, 35K50

1. Introduction. During the last two decades relevant progress has been made in the theory of boundary control of partial differential equations. In the case of equations of parabolic type, both variational and semigroup methods have been successfully applied (see, for instance, [L2], [DS], [S1], [S2], [Fa], [B1], [La], [LT1], [LT2], [F1], [F2]). Most of these (and other) papers deal with autonomous parabolic equations. Only [L2] and [DS] present results on the boundary control in the nonautonomous case, by variational techniques.

This paper concerns nonautonomous systems of parabolic type, from the point of view of semigroup theory. Our first purpose is to develop a suitable approach to nonhomogeneous initial boundary value problems based on the theory of evolution operators, in view of its application to boundary control problems.

Section 2 is devoted to this basic question. As in the autonomous case we are able to deal with control functions which are only square integrable in time and space. In particular our main goal is to derive a representation formula for solutions, similar to the classical one (see [B2], [La], [LT1]), which will prove to be very useful in the treatment of control problems.

Section 2 is organized as follows. Section 2.1 contains a detailed analysis of two concrete systems of equations of parabolic type with nonhomogeneous Dirichlet or Neumann boundary conditions, which motivate the abstract model to be introduced afterwards. In §§ 2.2 and 2.3 we study an abstract homogeneous nonautonomous parabolic Cauchy problem by the methods of [AT1], [AT2], which allow us to handle operators with variable domains and whose coefficients are just Hölder continuous in time. In § 2.4, by using the properties of the Dirichlet and Neumann maps, we obtain an abstract formulation of the concrete nonhomogeneous problems analyzed in § 2.1. Finally, in § 2.5 we derive the representation formula for solutions of the abstract version of nonhomogeneous initial boundary value problems; this formula is meaningful for nonregular boundary data and will be considered as the state equation for the control problems of § 3.

* Received by the editors November 28, 1988; accepted for publication (in revised form) November 17, 1989.

† Dipartimento di Metodi e Modelli Matematici per le Scienze Applicate, Università di Roma "La Sapienza," 00161 Roma, Italy.

‡ Dipartimento di Matematica, Università di Torino, 10123 Torino, Italy.

§ Dipartimento di Matematica "F. Enriques," Università di Milano, 20133 Milano, Italy.

The second part of the paper, namely § 3, deals with the linear-quadratic regulator (L-Q-R) problem, over finite time horizon, for an abstract evolution equation which includes the concrete models discussed in § 2. Here we follow the approach of [F2], based on a direct solution of the Riccati equation arising in the L-Q-R problem. However our assumptions on the final state cost operator P_T (see (3.13) below) are weaker than those imposed in [F2], and suggested by the more general results of [DI1]. We are able to solve directly the basic integral Riccati equation under general assumptions, much weaker and more natural for the applications than those imposed in [DS], where the Riccati equation was deduced from the optimality system (see Remark 2.3(iii) below). It turns out that our approach in the more general setting of nonautonomous problems still allows us to employ standard techniques of control theory. This fact lets us hope that many other results on boundary control problems, such as, e.g., infinite horizon optimal control [F2], [LT3], [DI2], [D], and the control of stochastic systems [F2], [I], can also be extended to the nonautonomous framework.

We conclude this section by listing some notation.

If X is a Banach space and $a<b$ we set:

$L^p(a,b;X):=$ space of strongly measurable functions $f:]a,b[\to X$ such that $\int_a^b \|f(t)\|_X^p\,dt<\infty$ ($1\leqq p<\infty$; obvious modifications for $p=\infty$); $C^k([a,b],X):=$ space of functions $f:[a,b]\to X$ which are k times continuously differentiable ($k\in\mathbb{N}$);

$C^{k+\vartheta}([a,b],X):=$ space of functions $f\in C^k([a,b],X)$ such that $f^{(k)}$ is ϑ-Hölder continuous ($k\in\mathbb{N}$, $\theta\in]0,1[$).

If X, Y are Banach spaces, we set:

$\mathscr{L}(X,Y):=$ space of bounded linear operators $T:X\to Y$;

$\mathscr{L}(X):=\mathscr{L}(X,X)$;

$C_s([a,b],\mathscr{L}(X,Y)):=$ space of operator-valued functions $T(\cdot):[a,b]\to\mathscr{L}(X,Y)$ which are strongly continuous, i.e., $T(\cdot)x\in C^0([a,b],Y)$ for each $x\in X$.

If H is a Hilbert space, we set:

$\Sigma(H):=$ space of self-adjoint operators $T\in\mathscr{L}(H)$;

$\Sigma^+(H):=$ space of self-adjoint operators $T\in\mathscr{L}(H)$ which are positive, i.e., $(Tx|x)_H\geqq 0$ for each $x\in H$.

If H is a Hilbert space and T is a linear operator in H, we set:

$D_T:=$ domain of T;

$\sigma(T):=$ spectrum of T;

$\rho(T):=$ resolvent set of T;

$T^*:=$ adjoint operator of T (whenever it exists).

Finally, if $m\in\mathbb{N}^*$ and Ω is a bounded open set of $\mathbb{R}^n$, we shall use the following spaces of $\mathbb{C}^m$-valued functions:

$$[C^k(\bar{\Omega})]^m,[C^{k+\vartheta}(\bar{\Omega})]^m,[L^p(\Omega)]^m(k\in\mathbb{N},\vartheta\in]0,1[,p\in[1,\infty]),$$

whose definitions are clear, and the usual Sobolev spaces

$[W^{\vartheta,p}(\Omega)]^m$, $[W^{\vartheta,p}(\partial\Omega)]^m(p\in[1,\infty[,\vartheta\in\mathbb{R})$,

$[W_0^{\vartheta,p}(\Omega)]^m(p\in[1,\infty[,\vartheta\in]1/p,\infty[)$.

2. Nonautonomous parabolic systems.

2.1. Two classical examples. We consider in this section two particular types of parabolic initial boundary value problems, namely, two parabolic systems with Dirichlet and Neumann conditions, respectively. We think of them as prototypes of the class of problems which are covered by the general theory of this section.

Let Ω be a bounded open set of $\mathbb{R}^n$, with boundary $\partial\Omega$ of class C^2. Fix $T>0$ and let $\{A_{sj}(t,x)\}_{s,j=1,\cdots,n}$ a set of $N\times N$ complex-valued matrices defined in $[0,T]\times\bar{\Omega}$, fulfilling the following hypotheses:

(2.1) (regularity)

$$A_{sj}\in C^{\alpha+1/2}([0,T],[C^0(\bar\Omega)]^{N^2})\cap C^{\alpha}([0,T],[C^1(\bar\Omega)]^{N^2}),$$

(2.2) (strong ellipticity)

$$\operatorname{Re}\sum_{sj=1}^{n}(A_{sj}(t,x)\cdot\eta_j\,|\,\eta_s)_{\mathbb{C}^N}$$

$$\geqq \nu\sum_{s=1}^{n}|\eta_s|^2\quad \forall\eta_1,\cdots,\eta_n\in\mathbb{C}^N,\quad \forall(t,x)\in[0,T]\times\bar\Omega\quad(\nu>0).$$

Under the above assumptions we consider the following problems:

$$(2.3)\quad \begin{aligned}&D_ty(t,x)-\sum_{sj=1}^{n}D_s[A_{sj}(t,x)\cdot D_jy(t,x)]+y(t,x)=0\quad\text{in }[0,T]\times\bar\Omega,\\&y(t,x)=u(t,x)\quad\text{in }[0,T]\times\partial\Omega,\\&y(0,x)=y_0(x)\quad\text{in }\bar\Omega;\end{aligned}$$

$$(2.4)\quad \begin{aligned}&D_ty(t,x)-\sum_{sj=1}^{n}D_s[A_{sj}(t,x)\cdot D_jy(t,x)]+y(t,x)=0\quad\text{in }[0,T]\times\bar\Omega,\\&\sum_{sj=1}^{n}A_{sj}(t,x)\cdot D_jy(t,x)\nu_s(x)=u(t,x)\quad\text{in }[0,T]\times\partial\Omega,\\&y(0,x)=y_0(x)\quad\text{in }\bar\Omega,\end{aligned}$$

where y_0, u are prescribed data on the parabolic boundary of $[0,T]\times\bar\Omega$. Here $\nu(x)$ is the unit outward normal vector at $x\in\partial\Omega$. It is well known that if u, y_0 are sufficiently smooth and fulfill suitable compatibility conditions at $\partial\Omega$ at $t=0$, then problems (2.3), (2.4) possess a unique solution; in addition we want to prove a representation formula for such solutions which will allow us to generalize the concept of solution to the case of less regular data u, y_0.

Concerning existence and uniqueness, we will invoke the results of Theorem 4.7 of [AT3]; to this purpose we just need that problems (2.3), (2.4) obey the requirements given there, namely, we need that the operator $\{\sum_{sj=1}^{n}D_s(A_{sj}(t,x)\cdot D_j)\}$, with boundary conditions of Dirichlet or of conormal derivative type, satisfies the ellipticity assumptions of [ADN] and [GG]. This is in fact true, as pointed out in Remark 2.3(i) below. Hence we can state the following propositions.

PROPOSITION 2.1. *Under assumptions* (2.1), (2.2), *let* $y_0\in[W^{2,2}(\Omega)]^N$, *and let u be the trace on* $[0,T]\times\partial\Omega$ *of a function* $U\in C^{\alpha}([0,T],[W^{1,2}(\Omega)]^N)\cap C^{\alpha+1/2}([0,T],[L^2(\Omega)]^N)$; *assume moreover that*

$$(2.5)\qquad \sum_{sj=1}^{n}A_{sj}(0,x)\cdot D_jy_0(x)\nu_s(x)=u(0,x)\quad \textit{a.e. on }\partial\Omega.$$

Then problem (2.4) *has a unique solution y such that*

$$(2.6)\qquad y\in C^1([0,T],[L^2(\Omega)]^N)\cap C([0,T],[W^{2,2}(\Omega)]^N).$$

PROPOSITION 2.2. *Under assumptions* (2.1), (2.2) *let* $y_0\in[W^{2,2}(\Omega)]^N$, *and let u be the trace on* $[0,T]\times\partial\Omega$ *of a function* $U\in C^{\alpha}([0,T],[W^{2,2}(\Omega)]^N)\cap C^{\alpha+1}([0,T],[L^2(\Omega)]^N)$; *assume moreover that*

$$(2.7)\qquad y_0(x)=u(0,x)\quad\textit{a.e. on }\partial\Omega.$$

Then problem (2.3) *has a unique solution y such that* (2.6) *holds.*

Proof. The proofs of Propositions 2.1, 2.2 follow by Theorem 4.7 of [AT3] with minor modifications (since the operators considered there are not in divergence form). □

Remark 2.3. (i) If we confine ourselves to problem (2.3), we may replace hypothesis (2.2) by the weaker one

$$\text{(2.8)}\quad \operatorname{Re}\sum_{sj=1}^{n}(A_{sj}(t,x)\xi_s\xi_j\cdot\eta\,|\,\eta)_{\mathbb{C}^N}\geqq\nu|\xi|^2|\eta|^2\quad\forall\xi\in\mathbb{R}^n,\quad\forall\eta\in\mathbb{C}^N,\ \forall(t,x)\in[0,T]\times\bar{\Omega};$$

then the operator $\{\sum_{sj=1}^{n}D_s(A_{sj}(t,x)\cdot D_j)\}$, with Dirichlet boundary conditions, still satisfies the ellipticity assumptions of [ADN] and [GG], as pointed out in [Am, pp. 659-660]. On the other hand, we are not able to prove the same assertion in the case of problem (2.4); that is, in order that the above operator, endowed with boundary conditions of conormal derivative type, satisfies the ellipticity assumptions of [ADN] and [GG], we need the stronger hypothesis (2.2) (this can be seen by adapting the argument of [ADN, p. 44]).

(ii) Adding lower order terms in problem (2.3), or (2.4), does not alter the situation: indeed, the change of unknown $v:=e^{\omega t}y$ (for a suitable $\omega\in\mathbb{R}$) leads to a new problem where the new differential operators still enjoy the properties stated in Proposition 2.4 below; in particular, the abstract hypothesis (2.29) is preserved.

(iii) In (2.1) it is assumed that the coefficients of the differential operators satisfy suitable Hölder conditions with respect to time. Such a requirement is necessary in order to fulfill the abstract assumption (2.29)(ii) below, which in turn allows us to construct the evolution operator for the abstract problem (2.28), with its regularity properties (3.4). If the coefficients are just bounded and measurable in t, then we can get some results for the concrete problems (2.3), (2.4) (see [LM1]), i.e., for the state equation; however the subsequent step, namely the study of the Riccati equation, seems very difficult and needs stronger hypotheses (see [DS]).

Existence and uniqueness of the solution of problems (2.3) and (2.4) is now guaranteed, at least for smooth data y_0, u. Our next goal is to establish a representation formula for the solution, which should possess the following features:

(i) It reduces to known representation formulas whenever they hold: see, e.g., [Te] for the autonomous versions of (2.3)-(2.4), [AT1] and [AT2] in the case of homogeneous boundary conditions, [B2] and [La] within the context of control theory;

(ii) It yields "weak" solutions, in some sense, when the data are less smooth;

(iii) It is handy from the point of view of control theory.

In order to construct such a formula, we need to reformulate problems (2.3), (2.4) in an abstract form, and to establish some properties of the evolution operators of the new problem. This will be the object of the next section.

2.2. The abstract formulation of initial boundary value problems. Consider again the situation of § 2.1, under assumptions (2.1), (2.2). If we define, for each $t\in[0,T]$, the differential operators

$$\text{(2.9)}\qquad \mathscr{A}(t,x,D)v:=\sum_{sj=1}^{n}D_s[A_{sj}(t,x)\cdot D_jv]-v,\qquad x\in\bar{\Omega},$$

$$\text{(2.10)}\qquad \mathscr{B}_0v:=v_{|\partial\Omega},$$

$$\text{(2.11)}\qquad \mathscr{B}_1(t,x,D):=\sum_{sj=1}^{n}A_{sj}(t,x)\nu_s(x)\cdot D_jv,\qquad x\in\partial\Omega,$$

then we can introduce the following linear operators:

$$\begin{aligned} &D_{A_0(t)} := \{v \in [W^{2,2}(\Omega)]^N : \mathscr{B}_0 v = 0\} = [W^{2,2}(\Omega) \cap W_0^{1,2}(\Omega)]^N, \\ &A_0(t)v := \mathscr{A}(t, \cdot, D)v, \end{aligned} \tag{2.12}$$

$$\begin{aligned} &D_{A_1(t)} := \{v \in [W^{2,2}(\Omega)]^N : \mathscr{B}_1(t, \cdot, D)v = 0\}, \\ &A_1(t)v := \mathscr{A}(t, \cdot, D)v, \end{aligned} \tag{2.13}$$

where $t \in [0, T]$.

The main properties of the operators $A_0(t)$, $A_1(t)$ are listed in the following proposition.

PROPOSITION 2.4. *Under assumptions* (2.1), (2.2) *we have for* $r = 0, 1$:

(i) *For each* $t \in [0, T]$, $A_r(t)$ *is the infinitesimal generator of an analytic semigroup in* $[L^2(\Omega)]^N$;

(ii) *for each* $t \in [0, T]$, $D_{A_r(t)}$ *is dense in* $[L^2(\Omega)]^N$;

(iii) *the family* $\{A_r(t)\}_{t \in [0,T]}$ *satisfies Hypothesis* II *of* [AT1], *i.e., there exists* $\vartheta_0 \in]\pi/2, \pi]$ *such that*

$$\begin{aligned} &\|A_r(t)[\lambda - A_r(t)]^{-1}[A_r(t)^{-1} - A_r(\tau)^{-1}]\|_{\mathscr{L}([L^2(\Omega)]^N)} \\ &\leqq c\frac{|t-\tau|^{\alpha+1/2}}{|\lambda|^{1/2}} \quad \forall t, \tau \in [0, T], \end{aligned} \tag{2.14}$$

provided λ *belongs to the sector* $S_{\vartheta_0} := \{z \in \mathbb{C} : |\arg z| < \vartheta_0\}$.

Proof. (i) It is well known (see [Am], [GG]) that the resolvent set of the operators $A_r(t)$ contains the sector

$$S_{\vartheta_0} + \omega = \{z \in \mathbb{C} : |\arg (z - \omega)| < \vartheta_0\}$$

for suitable $\vartheta_0 \in]\pi/2, \pi]$ and $\omega \in \mathbb{R}$; we want to show here that we can choose $\omega = 0$ and that

$$\|[\lambda - A_r(t)]^{-1}\|_{\mathscr{L}([L^2(\Omega)]^N)} \leqq \frac{c}{1+|\lambda|} \quad \forall \lambda \in \overline{S_{\vartheta_0}}. \tag{2.15}$$

Suppose first $r = 0$. Fix $t \in [0, T]$ and let $\lambda \in \mathbb{C}$ be such that either $\operatorname{Re} \lambda > 0$ or $(M/\nu)|\operatorname{Re} \lambda| \leqq \frac{1}{2}|\operatorname{Im} \lambda|$. For $v \in [W^{2,2}(\Omega) \cap W_0^{1,2}(\Omega)]^N$ set

$$f(x) := \lambda v(x) - \mathscr{A}(t, x, D)v, \qquad x \in \Omega.$$

Multiplying by v (with respect to the inner product of $[L^2(\Omega)]^N$) and integrating by parts, we get

$$\begin{aligned} &(1+\lambda)\int_\Omega |v|^2\,dx + \int_\Omega \sum_{sj=1}^{n} (A_{sj}(t, x) \cdot D_j v(x) \,|\, D_s v(x))_{\mathbb{C}^N}\,dx \\ &= \int_\Omega (f(x)\,|\,g(x))_{\mathbb{C}^N}\,dx. \end{aligned} \tag{2.16}$$

By taking the real part, we obtain by (2.2)

$$(1+\operatorname{Re}\lambda)\int_\Omega |v|^2\,dx + \nu\int_\Omega |Dv|^2\,dx \leqq \|f\|_{[L^2(\Omega)]^N} \cdot \|v\|_{[L^2(\Omega)]^N}; \tag{2.17}$$

on the other hand, by taking the imaginary part in (2.16),

$$|\mathrm{Im}\,\lambda|\int_\Omega |v|^2\,dx \leqq \|f\|_{[L^2(\Omega)]^N}\cdot\|v\|_{[L^2(\Omega)]^N}+M\int_\Omega |Dv|^2\,dx,$$

where

$$M:=\sum_{sj=1}^{n}\sup_{[0,T]\times\bar\Omega}|A_{sj}(t,x)|. \tag{2.18}$$

Hence by (2.17)

$$|\mathrm{Im}\,\lambda|\int_\Omega |v|^2\,dx \leqq \left(1+\frac{M}{\nu}\right)\|f\|_{[L^2(\Omega)]^N}\cdot\|v\|_{[L^2(\Omega)]^N}+\frac{M}{\nu}|\mathrm{Re}\,\lambda|\int_\Omega |Dv|^2\,dx.$$

Consequently, if $(M/\nu)|\mathrm{Re}\,\lambda|\leqq\frac{1}{2}|\mathrm{Im}\,\lambda|$ we deduce that

$$|\mathrm{Im}\,\lambda|\int_\Omega |v|^2\,dx \leqq 2\left(1+\frac{M}{\nu}\right)\|f\|_{[L^2(\Omega)]^N}\cdot\|v\|_{[L^2(\Omega)]^N}, \tag{2.19}$$

whereas if $(M/\nu)\,\mathrm{Re}\,\lambda>\frac{1}{2}|\mathrm{Im}\,\lambda|$ (2.17) yields

$$(1+\mathrm{Re}\,\lambda)\int_\Omega |v|^2\,dx \leqq \|f\|_{[L^2(\Omega)]^N}\cdot\|v\|_{[L^2(\Omega)]^N}. \tag{2.20}$$

Combining (2.19) and (2.20) we get the estimate

$$\|v\|_{[L^2(\Omega)]^N}\leqq\frac{c}{1+|\lambda|}\|\lambda v-\mathscr{A}(t,\cdot,D)v\|_{[L^2(\Omega)]^N}\quad\forall\lambda\in\overline{S_{\vartheta_0}},$$

where

$$\vartheta_0=\pi-\mathrm{arctg}\frac{2M}{\nu},\qquad c=2\left(1+\frac{M}{\nu}\right)\sqrt{1+(\nu/2M)^2}+1;$$

since we already know that $\rho(A_0(t))$ is not empty, the desired estimate (2.15) for $r=0$ follows from the above inequality by standard arguments.

The case $r=1$ is completely analogous and we find the same constants ϑ_0 and c.

The proof of part (ii) is obvious in both cases $r=0$, 1.

(iii) Consider the case $r=0$. Fix $f\in[L^2(\Omega)]^N$, and set $v:=[A_0(\tau)]^{-1}f$, $w:=[\lambda-A_0(t)]^{-1}[\lambda-A_0(\tau)]v$; then we must estimate

$$v-w=A_0(t)[\lambda-A_0(t)]^{-1}[A_0(t)^{-1}-A_0(\tau)^{-1}]f.$$

The function $v-w$ solves the problem

$$\lambda(v-w)-\mathscr{A}(t,\cdot,D)(v-w)=\sum_{sj=1}^{n}D_s([A_{sj}(t,\cdot)-A_{sj}(\tau,\cdot)]\cdot D_jv)\quad\text{in }\Omega,$$
$$v-w\in[W^{2,2}(\Omega)\cap W_0^{1,2}(\Omega)]^N.$$

Multiplying by $v-w$ (in $[L^2(\Omega)]^N$), an integration by parts yields

$$(1+\lambda)\int_\Omega |v-w|^2\,dx+\int_\Omega\sum_{sj=1}^{n}(A_{sj}(t,x)\cdot D_j(v-w)\,|\,D_s(v-w))_{\mathbb{C}^N}\,dx$$
$$=-\int_\Omega\sum_{sj=1}^{n}([A_{sj}(t,x)-A_{sj}(\tau,x)]\cdot D_jv\,|\,D_s(v-w))_{\mathbb{C}^N}\,dx,$$

which implies

$$\begin{aligned}(2.21)\qquad &(1+\operatorname{Re}\lambda)\int_\Omega |v-w|^2\,dx+\frac{\nu}{2}\int_\Omega |D(v-w)|^2\,dx\\ &\leqq\frac{1}{2\nu}\int_\Omega\sum_{sj=1}^n |A_{sj}(t,\cdot)-A_{sj}(\tau,\cdot)|^2|Dv|^2\,dx,\end{aligned}$$

$$\begin{aligned}(2.22)\qquad &|\operatorname{Im}\lambda|\int_\Omega |v-w|^2\,dx\leqq\frac{1}{2}\int_\Omega\sum_{sj=1}^n |A_{sj}(t,\cdot)-A_{sj}(\tau,\cdot)|^2|Dv|^2\,dx\\ &+\left(M+\frac{1}{2}\right)\int_\Omega |D(v-w)|^2\,dx.\end{aligned}$$

If we set

$$(2.23)\qquad N:=\sum_{sj=1}^n\ \sup_{0\leqq\tau<t\leqq T}\ \sup_{x\in\Omega}\frac{|A_{sj}(t,x)-A_{sj}(\tau,x)|}{|t-\tau|^{\alpha+1/2}},$$

then by (2.22) and (2.21) we easily get

$$\begin{aligned}|\operatorname{Im}\lambda|\int_\Omega |v-w|^2\,dx\leqq N^2\left[\frac{1}{2}+\left(M+\frac{1}{2}\right)\nu^{-2}\right]|t-\tau|^{2\alpha+1}\int_\Omega |Dv|^2\,dx\\ +\frac{2}{\nu}\left(M+\frac{1}{2}\right)|\operatorname{Re}\lambda|\int_\Omega |v-w|^2\,dx.\end{aligned}$$

Hence if $(2/\nu)(M+\frac{1}{2})|\operatorname{Re}\lambda|\leqq\frac{1}{2}|\operatorname{Im}\lambda|$

$$(2.24)\qquad |\operatorname{Im}\lambda|\int_\Omega |v-w|^2\,dx\leqq N^2\left[1+\frac{2M+1}{\nu^2}\right]|t-\tau|^{2\alpha+1}\int_\Omega |Dv|^2\,dx,$$

whereas if $(2/\nu)(M+\frac{1}{2})\operatorname{Re}\lambda>\frac{1}{2}|\operatorname{Im}\lambda|$

$$(2.25)\qquad \operatorname{Re}\lambda\int_\Omega |v-w|^2\,dx\leqq\frac{N^2}{2\nu}|t-\tau|^{2\alpha+1}\int_\Omega |Dv|^2\,dx.$$

Recalling that, by (2.17),

$$\int_\Omega |Dv|^2\,dx\leqq\frac{1}{\nu}\int_\Omega |f|^2\,dx,$$

we conclude that

$$|\lambda|\int_\Omega |v-w|^2\,dx\leqq\frac{N^2}{\nu}\left(1+\frac{2M+1}{\nu^2}\right)\sqrt{1+[\nu/(4M+2)]^2}\,|t-\tau|^{2\alpha+1}\int_\Omega |f|^2\,dx,$$

and (2.14) follows for $r=0$, with

$$\vartheta_0=\pi-\operatorname{arctg}\frac{4M+2}{\nu},\qquad c=N\nu^{-1/2}\left(1+\frac{2M+1}{\nu^2}\right)^{1/2}\left[1+\left(\frac{\nu}{4M+2}\right)^2\right]^{1/4}.$$

Concerning the case $r=1$, we proceed similarly and we find that $v-w$ now solves the problem

$$\lambda(v-w)-\mathscr{A}(t,\cdot,D)(v-w)=\sum_{sj=1}^n D_s([A_{sj}(t,\cdot)-A_{sj}(\tau,\cdot)]\cdot D_jv)\quad\text{in }\Omega,$$

$$\sum_{sj=1}^n A_{sj}(t,\cdot)\nu_s\cdot D_j(v-w)=\sum_{sj=1}^n [A_{sj}(\tau,\cdot)-A_{sj}(t,\cdot)]\nu_s\cdot D_jv\quad\text{on }\partial\Omega;$$

arguing as above, and taking into account the boundary conditions, we obtain the result with the same constants ϑ_0 and c. The proof of Proposition 2.4 is complete. □

Remark 2.5. The estimates (2.15) and (2.14) do not need that A_{sj} belongs to $C^\alpha([0, T], [C^1(\bar{\Omega})]^N)$.

Consider now the operators $A_r(t)^*$, i.e., the adjoint operators of $A_r(t)(t \in [0, T], r = 0, 1)$. It is easy to verify that they are defined by

$$\begin{aligned} &D_{A_0(t)^*} := [W^{1,2}(\Omega) \cap W_0^{1,2}(\Omega)]^N, \\ (2.26)\qquad &A_0(t)^* y := \overline{\mathscr{A}(t, \cdot, D)} y = \sum_{sj=1}^{n} D_j[\overline{{}^tA_{sj}(t, \cdot)} \cdot D_s y] - y, \end{aligned}$$

$$\begin{aligned} &D_{A_1(t)^*} := \left\{ y \in [W^{2,2}(\Omega)]^N : \overline{\mathscr{B}_1(t, \cdot, D)} y = \sum_{sj=1}^{n} \overline{{}^tA_{sj}(t, \cdot)} \nu_s \cdot D_j y = 0 \right\}, \\ (2.27)\qquad &A_1(t)^* y := \overline{\mathscr{A}(t, \cdot, D)} y, \end{aligned}$$

where $\overline{{}^tA_{sj}}$ is the matrix whose elements are the conjugates of the elements of the transposed ${}^tA_{sj}$ of A_{sj}. Consequently, it is clear that the following result holds.

PROPOSITION 2.6. *All statements of Proposition* 2.4 *are true if* $A_r(t)$ *is replaced by* $A_r(t)^*$.

The results of Propositions 2.4 and 2.6 allow us to apply to the operators $\{A_r(t)\}_{t \in [0,T]}$, $\{A_r(t)^*\}_{t \in [0, T]}$ the abstract theory of [AT1], [AT2], and [Ac] concerning linear nonautonomous parabolic Cauchy problems of the following kind:

$$(2.28)\qquad \begin{aligned} &u'(t) - A(t)u(t) = f(t), \qquad t \in [0, T], \\ &u(0) = x, \end{aligned}$$

where $f \in C([0, T], E)$, $x \in E$ (E being a general Banach space) and $\{A(t)\}_{t \in [0,T]}$ fulfills (2.15) and (2.14). In the next section we will recall some facts concerning a problem such as (2.28).

2.3. The study of the abstract problem. We now consider problem (2.28), but we restrict our considerations to the case of a Hilbert space H, which is enough for our successive applications. We assume that:

(2.29) $\{A(t)\}_{t \in [0,T]}$ is a family of closed linear operators in H, such that:

$$\text{(i)}\quad \|[\lambda - A(t)]^{-1}\|_{\mathscr{L}(H)} \leqq \frac{M}{1 + |\lambda|} \quad \forall \lambda \in \overline{S_{\vartheta_0}}, \quad \forall t \in [0, T],$$

$$\text{(ii)}\quad \|A(t)[\lambda - A(t)]^{-1}[A(t)^{-1} - A(s)^{-1}]\|_{\mathscr{L}(H)} \leqq B \frac{|t - s|^{\alpha + 1/2}}{|\lambda|^{1/2}}$$

$$\forall \lambda \in S_{\vartheta_0}, \quad \forall t, s \in [0, T],$$

where $\vartheta_0 \in]\pi/2, \pi]$ and α, M, $B > 0$.

In particular, $A(t)$ generates an analytic semigroup $e^{\xi A(t)}$ which can be represented as a Dunford integral:

$$(2.30)\qquad e^{\xi A(t)} = (2\pi i)^{-1} \int_\Gamma e^{\xi \lambda} [\lambda - A(t)]^{-1} \, d\lambda,$$

Γ being a smooth path contained in S_{ϑ_0} and joining $+\infty\, e^{-i\vartheta}$ to $+\infty\, e^{i\vartheta}$, $\vartheta \in]\pi/2, \vartheta_0[$. Moreover, the fractional powers $[-A(t)]^{\gamma}$ are well defined and we have the representations

$$(2.31) \qquad [-A(t)]^{-\gamma} = (2\pi i)^{-1} \int_{\Gamma'} (-\lambda)^{-\gamma} [\lambda - A(t)]^{-1}\, d\lambda, \qquad \gamma > 0,$$

$$(2.32) \qquad [-A(t)]^{-\gamma}\, e^{\xi A(t)} = (2\pi i)^{-1} \int_{\Gamma'} (-\lambda)^{-\gamma}\, e^{\xi\lambda} [\lambda - A(t)]^{-1}\, d\lambda$$

where $\Gamma' \subset S_{\theta_0}$ joins $+\infty\, \mathrm{e}^{-i\theta}$ to $+\infty\, \mathrm{e}^{i\theta}$ leaving 0 on its right-hand side. We also recall the well-known continuous inclusions

$$(2.33) \qquad \begin{gathered} D_{A(t)}(\gamma + \varepsilon, \infty) \subset D_{[-A(t)]^{\gamma}} \subset D_{A(t)}(\gamma, \infty) \\ \forall \gamma \in]0, 1[, \quad \forall \varepsilon \in]0, 1-\gamma[, \quad \forall t \in [0, T]; \end{gathered}$$

here $D_{A(t)}(\gamma, p)$, $1 \leqq p \leqq \infty$, is the real interpolation space $(D_{A(t)}, H)_{1-\gamma,\infty}$ introduced in [LP], which can be characterized in the following way:

$$(2.34) \qquad D_{A(t)}(\gamma, p) = \{x \in H\colon \xi \to \xi^{-\gamma} \| [e^{\xi A(t)} - 1]x \|_H \in L^p(0, \infty; d\xi/\xi)\}.$$

We need the following lemma.

LEMMA 2.7. *Under assumption* (2.29) *we have for each* t, $s \in [0, T]$:

(i) $$\| [-A(t)]^{-\vartheta} - [-A(s)]^{-\vartheta} \|_{\mathscr{L}(H)} \leqq \begin{cases} c(\vartheta, \alpha) |t-s|^{\alpha+1/2} & \textit{if } \vartheta > \tfrac{1}{2}, \\ c(\vartheta, \alpha, \sigma) |t-s|^{\sigma(\alpha+1/2)} \quad \forall \sigma \in]0, 2\vartheta[& \textit{if } \vartheta \in]0, \tfrac{1}{2}]; \end{cases}$$

(ii) $$\| [-A(t)]^{\vartheta}\, e^{\xi A(t)} \|_{\mathscr{L}(H)} \leqq c(\vartheta) \xi^{-\vartheta} \quad \forall \xi > 0;$$

(iii) $$\| [-A(t)]^{\vartheta}\, e^{\xi A(t)} - [-A(s)]^{\vartheta}\, e^{\xi A(s)} \|_{\mathscr{L}(H)} \leqq c(\vartheta, \alpha) |t-s|^{\alpha+1/2} \xi^{-\vartheta-1/2} \quad \forall \xi > 0.$$

Proof. (i) An easy check shows that

$$[\lambda - A(t)]^{-1} - [\lambda - A(s)]^{-1} = -A(t)[\lambda - A(t)]^{-1} [A(t)^{-1} - A(s)^{-1}] A(s) [\lambda - A(s)]^{-1};$$

hence by (2.31) and (2.29) we get

$$\| [-A(t)]^{-\vartheta} - [-A(s)]^{-\vartheta} \|_{\mathscr{L}(H)} \leqq c \int_{\Gamma'} |\lambda|^{-\vartheta} \left[\frac{M}{1 + |\lambda|} \right]^{1-\sigma} \left[\frac{B|t-s|^{\alpha+1/2}}{|\lambda|^{1/2}} \right]^{\sigma} |d\lambda| \qquad \forall \sigma \in [0, 1],$$

which easily leads to the result.

Parts (ii) and (iii) follow similarly by (2.32) and (2.29). □

We are ready to state the main result concerning problem (2.28).

PROPOSITION 2.8. *Under assumption* (2.29), *the evolution operator* $U(t, s) \in \mathscr{L}(H, D_{A(t)})$, *associated to problem* (2.28), *exists and possesses the following properties*:

(i) $(t, s) \to U(t, s) \in C(\Delta, \mathscr{L}(H)) \cap C_s(\bar{\Delta}, \mathscr{L}(H))$, *where* $\Delta := \{(t, s) \in [0, T]\colon s < t\}$, *and*

$$U(t, t) = 1, \quad U(t, \tau) U(\tau, s) = U(t, s) \quad \forall \tau \in [s, t];$$

(ii) $(t, s) \to A(t) U(t, s) \in C(\Delta, \mathscr{L}(H))$ *and*

$$A(t) U(t, s) = \frac{\partial}{\partial t} U(t, s), \quad \| A(t) U(t, s) \|_{\mathscr{L}(H)} \leqq M_1 (t-s)^{-1} \quad \forall 0 \leqq s \leqq t \leqq T;$$

(iii) *If* $s \in [0, t[$ *and* $x \in D_{A(s)}$, *then* $\partial/\partial s U(t, s)x = -U(t, s)A(s)x$ *in the following sense*:

$$h^{-1}[U(t, s+h) - U(t, s)]x \to -U(t, s)A(s)x \quad \text{in } H \quad \text{as } h \to 0_+,$$

$$h^{-1}[U(t, s+h) - U(t, s)]A(s+h)^{-1}A(s)x \to -U(t, s)A(s)x \quad \text{in } H \quad \text{as } h \to 0_-;$$

(iv) *If* γ, $\beta \in [0, 1]$, *then* $(t, s) \to [-A(t)]^{\gamma} U(t, s)[-A(s)]^{-\beta} \in C(\Delta, \mathscr{L}(H))$ *and*

$$\|[-A(t)]^{\gamma} U(t, s)[-A(s)]^{-\beta}\|_{\mathscr{L}(H)} \leqq M_{\gamma\beta}[(t-s)^{\beta-\gamma} + 1] \quad \forall 0 \leqq s \leqq t \leqq T;$$

(v) *If* $0 \leqq \gamma \leqq \beta \leqq 1$, *then* $(t, s) \to [-A(t)]^{\gamma} U(t, s)[-A(s)]^{-\beta} \in C(\bar{\Delta}, \mathscr{L}(H))$.

Proof. Parts (i)-(iii) are proved in Theorem 3.2 of [Ac] (recall that the domains $D_{A(t)}$ are dense in H here), with the exception of the assertion $(t, s) \to U(t, s) \in C_s(\bar{\Delta}, \mathscr{L}(H))$. In order to show this property, we first recall that by the density of domains and by Lemma 1.9(i) of [AT1] we have

$$\lim_{n\to\infty} \|x - n[n - A(\tau)]^{-1}x\|_H = 0 \quad \forall x \in H, \quad \forall \tau \in [0, T], \tag{2.35}$$

$$\begin{aligned} &\|A(s)[n - A(s)]^{-1} - A(\tau)[n - A(\tau)]^{-1}\|_{\mathscr{L}(H)} \leqq Bn^{1/2}|\tau - s|^{\alpha+1/2} \\ &\qquad\qquad \forall n \in \mathbb{N}^+, \quad \forall \tau, s \in [0, T]. \end{aligned} \tag{2.36}$$

Now let $(\tau, \tau) \in \partial\Delta$, $x \in H$; then we have (see [Ac, formula (2.6)]):

$$\begin{aligned} U(t, s)x - x &= [e^{(t-s)A(s)}x - x] + \int_s^t Z(r, s)x\, dr \\ &= [e^{(t-s)A(s)} - 1]\{[x - n[n - A(\tau)]^{-1}x] \\ &\quad + [n[n - A(\tau)]^{-1} - n[n - A(s)]^{-1}]x\} \\ &\quad + \int_0^{t-s} e^{\sigma A(s)}[nA(s)[n - A(s)]^{-1} - nA(\tau)[n - A(\tau)]^{-1}]x\, d\sigma \\ &\quad + \int_0^{t-s} e^{\sigma A(s)} nA(\tau)[n - A(\tau)]^{-1}x\, d\sigma + \int_s^t Z(r, s)x\, dr; \end{aligned}$$

hence by (2.36) and Lemma 2.2(i) of [Ac] we easily obtain

$$\begin{aligned} \|U(t, s)x - x\|_H \leqq c(M, \beta, \alpha)\{&(1 + n(t-s))\|x - n[n - A(\tau)]^{-1}x\|_H \\ &+ \|x\|_H[(1 + n(t-s))n^{1/2}|\tau - s|^{\alpha+1/2} + (t-s)^{\alpha}]\}. \end{aligned}$$

By (2.35) there exists $\nu_\varepsilon \in \mathbb{N}^*$ such that

$$\|x - \nu_\varepsilon[\nu_\varepsilon - A(\tau)]^{-1}x\|_H < \tfrac{1}{2}\varepsilon^2[c(M, \beta, \alpha)]^{-1};$$

choosing $n = \nu_\varepsilon$ and $\delta_\varepsilon > 0$ such that

$$c(M, \beta, \alpha)[(1 + \nu_\varepsilon\delta_\varepsilon)\tfrac{1}{2}\varepsilon^2[c(M, \beta, \alpha)]^{-1} + \|x\|_H(1 + \nu_\varepsilon\delta_\varepsilon)\nu_\varepsilon^{1/2}\delta_\varepsilon^{\alpha+1/2} + \|x\|_H\delta_\varepsilon^{\alpha}] < \varepsilon,$$

we immediately get

$$\|U(t, s)x - x\|_H \leqq \varepsilon \quad \text{if } |t - \tau| + |\tau - s| < \delta_\varepsilon.$$

Note that, in particular, the above proof shows that

$$(t, s) \to e^{(t-s)A(s)} \in C_s(\bar{\Delta}, \mathscr{L}(H)). \tag{2.37}$$

Let us prove (iv). We write

$$[-A(t)]^{\gamma} U(t, s)[-A(s)]^{-\beta} = -[-A(t)]^{\gamma-1}[A(t)U(t, s)][-A(s)]^{-\beta};$$

since each operator in the right-hand side is in $C(\Delta, \mathscr{L}(H))$, we get that the left-hand side also belongs to $C(\Delta, \mathscr{L}(H))$. In order to prove the estimate, we remark that if $x \in H$, then $t \to U(t, s)[-A(s)]^{-\beta}x$ is the classical solution [AT1, Def. 1.6] of the problem

$$(2.38) \qquad \begin{aligned} &u'(t)-A(t)u(t)=0, \qquad t \in]s, T], \\ &u(0)=[-A(s)]^{-\beta}x, \end{aligned}$$

and consequently [AT1, Thm. 6.3(i)] $t \to [A(t)U(t, s)][-A(s)]^{-\beta}x$ solves the integral equation

$$(2.39) \qquad v(t)-[Q_s v](t)=A(t)\, e^{(t-s)A(t)}[-A(s)]^{-\beta}x, \qquad t \in [s, T],$$

where the integral operator Q_s is defined [Ac, (2.1)-(2.2)] by

$$(2.40) \qquad [Q_s v](t) := \int_s^t A(t)^2\, e^{(t-\tau)A(t)}[A(t)^{-1}-A(\tau)^{-1}]v(\tau)\, d\tau, \qquad t \in [s, T].$$

Hence we can write

$$\begin{aligned} &[-A(t)]^{\gamma}U(t, s)[-A(s)]^{-\beta}x \\ &\quad = -[-A(t)]^{\gamma-1}[Q_s([A(t)U(t, s)][-A(s)]^{-\beta}x)](t) \\ &\qquad -[-A(t)]^{\gamma}\, e^{(t-s)A(t)}[-A(s)]^{-\beta}x \\ (2.41) \quad &\quad = -\int_s^t [-A(t)]^{\gamma+1}\, e^{(t-\tau)A(t)}[A(t)^{-1}-A(\tau)^{-1}]A(\tau)U(\tau, s)[-A(s)]^{-\beta}x\, d\tau \\ &\qquad -[[-A(t)]^{\gamma}\, e^{(t-s)A(t)}-[-A(s)]^{\gamma}\, e^{(t-s)A(s)}][-A(s)]^{-\beta}x \\ &\qquad -[-A(s)]^{\gamma-\beta}\, e^{(t-s)A(s)}x, \end{aligned}$$

and by Lemma 2.7 we readily obtain the result.

Finally, we prove (v). By (2.41) it is enough to show that if $(t, s) \to (\tau, \tau)$ in Δ and $x \in H$, then

$$\|([-A(s)]^{\gamma-\beta}\, e^{(t-s)A(s)}-[-A(\tau)]^{\gamma-\beta})x\|_H \to 0.$$

If $\beta=\gamma$ this follows by (2.37); otherwise we can write

$$\begin{aligned} &([-A(s)]^{\gamma-\beta}\, e^{(t-s)A(s)}-[-A(t)]^{\gamma-\beta})x \\ &\quad = \int_0^{t-s} [-A(s)]^{\gamma-\beta+1}\, e^{\xi A(s)}x\, d\xi + [[-A(s)]^{\gamma-\beta}-[-A(\tau)]^{\gamma-\beta}]x, \end{aligned}$$

which by Lemma 2.7 implies the result. □

Assume now that the adjoint operator $A(t)^*$ of $A(t)$ also satisfies (2.29), i.e.,

$$(2.42) \qquad \begin{aligned} &\text{(i)} \quad \|[\lambda - A(t)^*]^{-1}\|_{\mathscr{L}(H)} \leqq \frac{M}{1+|\lambda|} \quad \forall \lambda \in \overline{S_{\vartheta_0}}, \quad \forall t \in [0, T], \\ &\text{(ii)} \quad \|A(t)^*[\lambda - A(t)^*]^{-1}[[A(t)^*]^{-1}-[A(s)^*]^{-1}]\|_{\mathscr{L}(H)} \\ &\qquad \leqq B\frac{|t-s|^{\alpha+1/2}}{|\lambda|^{1/2}} \quad \forall \lambda \in S_{\vartheta_0}, \quad \forall t, s \in [0, T]. \end{aligned}$$

Then Proposition 2.8 also holds for $A(t)^*$.

The next result concerns the adjoint operator $U(t, s)^*$ of the evolution operator $U(t, s)$ relative to $A(t)$.

PROPOSITION 2.9. *Under assumptions* (2.29), (2.42) *let* $U(t, s)$ *be the evolution operator of problem* (2.28). *Then*:

(i) $U(t, s)^* \in \mathscr{L}(H, D_{A(s)^*})$, *for all* $s \in [0, t[$;

(ii) *For each* $\varphi \in H$, $s \to U(t, s)^*\varphi$ *solves the problem*

$$\begin{aligned} &\frac{d}{ds} U(t, s)^*\varphi = -A(s)^* U(t, s)^*\varphi, \qquad s \in [0, t[, \\ &U(t, t)^*\varphi = \varphi. \end{aligned} \tag{2.43}$$

Proof. First of all, we show that the solution of (2.43) exists. Fix $t_0 \in]0, T]$ and set

$$V(t_0; t, s) := \text{the evolution operator relative to } B(t) := A(t_0 - t)^*, \quad t \in [0, t_0]. \tag{2.44}$$

This means that

$$\begin{aligned} &\frac{d}{dt} V(t_0; t, s)\varphi = A(t_0 - t)^* V(t_0; t, s)\varphi, \qquad t \in]s, t_0], \\ &V(t_0; s, s)\varphi = \varphi. \end{aligned} \tag{2.45}$$

Set $W(t, s) := V(t; t - s, 0)$, $s \in [0, t]$. Then, applying Proposition 2.8 to problem (2.45), we get $W(t, s) \in \mathscr{L}(H, D_{A(s)^*})$ and

$$\begin{aligned} \frac{d}{ds} W(t, s)\varphi = -\left[\frac{d}{d\tau} V(t; \tau, 0)\varphi\right]_{\tau = t-s} &= -[A(t-\tau)^* V(t; \tau, 0)\varphi]_{\tau = t-s} \\ &= -A(s)^* W(t, s)\varphi, \qquad s \in [0, t[, \end{aligned}$$

$$W(t, t)\varphi = V(t; 0, 0)\varphi = \varphi,$$

i.e., $W(t, s)$ solves (2.43). The proof will be complete by showing that

$$V(t; t - s, 0) = W(t, s) = U(t, s)^*. \tag{2.46}$$

Indeed for $r \in]s, t[$ we have

$$\begin{aligned} &\frac{d}{dr} (W(t, r)\varphi \mid U(r, s)x)_H \\ &\qquad = -(A(r)^* W(t, r)\varphi \mid U(r, s)\varphi)_H + (W(t, r)\varphi \mid A(r) U(r, s)\varphi)_H = 0, \end{aligned}$$

so that $(W(t, r)\varphi \mid U(r, s)x)_H = \text{const.}$ for all $r \in [s, t]$. As $r \to t^-$ and $r \to s^+$ we get

$$(\varphi \mid U(r, s)x)_H = (W(t, r)\varphi \mid x)_H \quad \forall \varphi, x \in H,$$

i.e., $W(t, s) = U(t, s)^*$. □

COROLLARY 2.10. *Under assumptions* (2.29), (2.42) *we have for* $\gamma, \beta \in [0, 1]$:

$$\|[-A(s)^*]^\gamma U(t, s)^* [-A(t)^*]^{-\beta}\|_{\mathscr{L}(H)} \leqq M_{\gamma\beta}[(t-s)^{\beta-\gamma} + 1] \quad \forall 0 \leqq s < t \leqq T.$$

Proof. We have by (2.44) and (2.46)

$$[-A(s)^*]^\gamma U(t, s)^* [-A(t)^*]^{-\beta} = [[-B(\tau)]^\gamma V(t; \tau, 0)[-B(0)]^{-\beta}]_{\tau = t-s};$$

hence the result follows by applying Proposition 2.8 to problem (2.45). □

COROLLARY 2.11. *Under assumptions* (2.29), (2.42) *let* $\beta, \gamma \in [0, 1]$. *Then for* $0 \leqq s < t \leqq T$ *the closed linear operator*

$$[-A(t)]^{-\beta} U(t, s)[-A(s)]^\gamma$$

possesses an extension $\overline{[-A(t)]^{-\beta}U(t,s)[-A(s)]^{\gamma}} \in \mathscr{L}(H)$, *which satisfies*

$$\|\overline{[-A(t)]^{-\beta}U(t,s)[-A(s)]^{\gamma}}\|_{\mathscr{L}(H)} \leqq M_{\gamma\beta}[(t-s)^{\beta-\gamma}+1] \quad \forall 0 \leqq s < t \leqq T.$$

Proof. As

$$[[-A(s)]^{\gamma}]^* = [[-A(s)]^*]^{\gamma} \quad \forall \gamma \in [0,1],$$

if $x \in D_{[-A(s)]^{\gamma}}$ and $\varphi \in H$ we have

$$([-A(t)]^{-\beta}U(t,s)[-A(s)]^{\gamma}x\,|\,\varphi)_H = (x\,|\,[-A(s)^*]^{\gamma}U(t,s)^*[-A(t)^*]^{-\beta}\varphi)_H,$$

and by Corollary 2.10

$$|([-A(t)]^{-\beta}U(t,s)[-A(s)]^{\gamma}x\,|\,\varphi)_H| \leqq M_{\gamma\beta}[(t-s)^{\beta-\gamma}+1]\|x\|_H\,\|\varphi\|_H;$$

choosing $y := [-A(t)]^{-\beta}U(t,s)[-A(s)]^{\gamma}x$ and $\varphi := y/\|y\|_H$, by the density of $D_{[-A(s)]^{\gamma}}$ in H we get the result. □

The study of the abstract problem (2.28) (which concerns homogeneous boundary conditions) is complete. In the next section we will introduce nonhomogeneous boundary data in the abstract framework.

2.4. The Dirichlet and Neumann maps. Let us go back to problems (2.3), (2.4): we will examine the regularity properties of the Dirichlet and Neumann maps $G_0(t)$, $G_1(t)$ which are defined by (see (2.9)-(2.11)):

$$u := G_0(t)g \Leftrightarrow \begin{cases} \mathscr{A}(t,\cdot,D)u = 0 & \text{in } \Omega, \\ \mathscr{B}_0 u = g & \text{on } \partial\Omega, \end{cases} \tag{2.47}$$

$$u := G_1(t)g \Leftrightarrow \begin{cases} \mathscr{A}(t,\cdot,D)u = 0 & \text{in } \Omega, \\ \mathscr{B}_1(t,\cdot,D)u = g & \text{on } \partial\Omega. \end{cases} \tag{2.48}$$

PROPOSITION 2.12. *Let* $A_0(t)$, $A_1(t)$ *be defined by* (2.12), (2.13), *respectively. If* $r = 0, 1$ *the operator* $G_r(t)$ *is well defined from* $[L^2(\partial\Omega)]^N$ *into* $D_{[-A_r(t)]^{\vartheta}}$, *for each* $\vartheta \in]0, \alpha_r[$, *where* $\alpha_0 := \frac{1}{4}$ *and* $\alpha_1 := \frac{3}{4}$. *Moreover,*

$$t \to [-A_r(t)]^{\vartheta}G_r(t) \in L^{\infty}(0,T;\mathscr{L}([L^2(\partial\Omega)]^N,[L^2(\Omega)]^N)) \quad \forall \vartheta \in]0,\alpha_r[.$$

Proof. This result was pointed out in [La] assuming $\partial\Omega \in C^{\infty}$; here we give an independent proof.

Let us start with the case $r = 0$. Fix $t \in [0,T]$, let $g \in [W^{1/2,2}(\partial\Omega)]^N$, and consider the variational problem corresponding to (2.9), (2.10):

$$\begin{aligned} &\mathscr{A}(t,\cdot,D)u_0 = 0 && \text{in } \Omega, \\ &u_0 = g && \text{on } \partial\Omega, \end{aligned} \tag{2.49}$$

which means

$$\begin{aligned} &\sum_{sj=1}^{n}\int_{\Omega}[(A_{sj}(t,x)\cdot D_j u_0\,|\,D_s\varphi)_{\mathbb{C}^N} + (u_0\,|\,\varphi)_{\mathbb{C}^N}]\,dx = 0 \quad \forall \varphi \in [C_0^{\infty}(\Omega)]^N, \\ &u_0 - G \in [W_0^{1,2}(\Omega)]^N, \end{aligned} \tag{2.50}$$

where G is an element of $[W^{1,2}(\Omega)]^N$ whose trace on $\partial\Omega$ is g, and such that

$$\|g\|_{[W^{1/2,2}(\partial\Omega)]^N} \leqq c_0\|G\|_{[W^{1,2}(\Omega)]^N} \leqq c_1\|g\|_{[W^{1/2,2}(\partial\Omega)]^N}. \tag{2.51}$$

By Poincaré inequality and Lax-Milgram theorem, problem (2.50) is uniquely solvable: we denote its solution u_0 by $S_0(t)g$, and we easily get the estimate

$$\|S_0(t)g\|_{[W^{1,2}(\Omega)]^N} \leqq c(M, \nu, c_0, c_1)\|g\|_{[W^{1/2,2}(\partial\Omega)]^N} \quad \forall g \in [W^{1/2,2}(\partial\Omega)]^N, \tag{2.52}$$

where $M := \sum_{sj=1}^n \|A_{sj}\|_{C([0,T],[C^1(\bar{\Omega})]^{N^2})}$.

Note that if $g \in [W^{3/2,2}(\partial\Omega)]^N$, then by (2.1) and the classical results of [ADN] we have $S_0(t) \in [W^{2,2}(\Omega)]^N$ and

$$\|S_0(t)g\|_{[W^{2,2}(\Omega)]^N} \leqq c\|g\|_{[W^{3/2,2}(\partial\Omega)]^N} \quad \forall g \in [W^{3/2,2}(\partial\Omega)]^N. \tag{2.53}$$

We want now to estimate $S_0(t)$ in a lower norm. For $g \in [W^{1/2,2}(\partial\Omega)]^N$ set $u_0 := S_0(t)g$ and let ψ be the variational solution of

$$\begin{aligned} &\sum_{sj=1}^n \int_\Omega [(\overline{{}^tA_{sj}(t,x)} \cdot D_s\psi | D_j\varphi)_{\mathbb{C}^N} + (\psi|\varphi)_{\mathbb{C}^N}]\, dx = \int_\Omega (u_0|\varphi)_{\mathbb{C}^N}\, dx \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall \varphi \in [C_0^\infty(\Omega)]^N, \\ &\psi = 0 \quad \text{on } \partial\Omega; \end{aligned} \tag{2.54}$$

as $u_0 \in [W^{1,2}(\Omega)]^N \subset [L^2(\Omega)]^N$, we have $\psi \in [W^{2,2}(\Omega)]^N \cap [W_0^{1,2}(\Omega)]^N$, and

$$-\sum_{sj=1}^n D_j(\overline{{}^tA_{sj}(t,\cdot)} \cdot D_s\psi) + \psi = u_0 \quad \text{a.e. in } \Omega; \tag{2.55}$$

in addition

$$\|\psi\|_{[W^{3/2,2}(\partial\Omega)]^N} \leqq c\|\psi\|_{[W^{2,2}(\Omega)]^N} \leqq c\|u_0\|_{[L^2(\Omega)]^N}. \tag{2.56}$$

By density we may choose $\varphi = \psi$ in (2.50); an integration by parts yields

$$\int_\Omega \left(u_0 \middle| \psi - \sum_{sj=1}^n D_j[\overline{{}^tA_{sj}(t,x)} \cdot D_s\psi]\right)_{\mathbb{C}^N} dx = -\int_{\partial\Omega} \left(u_0 \middle| \sum_{sj=1}^n \overline{{}^tA_{sj}(t,x)} \cdot D_s\psi\nu_j\right)_{\mathbb{C}^N} d\sigma,$$

and by (2.55) (since $u_0 = g$ on $\partial\Omega$)

$$\int_\Omega |u_0|^2\, dx = -\int_{\partial\Omega} \left(g \middle| \sum_{sj=1}^n \overline{{}^tA_{sj}(t,x)} \cdot D_s\psi\nu_j\right)_{\mathbb{C}^N} d\sigma. \tag{2.57}$$

Now, as $\partial\Omega \in C^2$, the function $d(x)$, i.e., the distance of $x \in \Omega$ from $\partial\Omega$, is of class C^2 in a neighbourhood of $\partial\Omega$ and $Dd(x) = -\nu(x)$ on $\partial\Omega$ (see [GT, Appendix]); moreover we can clearly modify $d(x)$ inside Ω in order to get $d \in C^2(\bar{\Omega})$. Hence by (2.57) and (2.56) it follows that

$$\begin{aligned} \|u_0\|^2_{[L^2(\Omega)]^N} &= |\langle g, \overline{{}^tA_{sj}(t,x)} \cdot D_s\psi\nu_j\rangle_{[W^{-1/2,2}(\partial\Omega)]^N,[W^{1/2,2}(\partial\Omega)]^N}| \\ &\leqq \|g\|_{[W^{-1/2,2}(\partial\Omega)]^N} \|\overline{{}^tA_{sj}} \cdot D_s\psi\nu_j\|_{[W^{1/2,2}(\partial\Omega)]^N} \\ &\leqq c\|g\|_{[W^{-1/2,2}(\partial\Omega)]^N} \|\overline{{}^tA_{sj}} \cdot D_s\psi D_jd\|_{[W^{1,2}(\Omega)]^N} \\ &\leqq c(M,\Omega)\|g\|_{[W^{-1/2,2}(\partial\Omega)]^N} \|\psi\|_{[W^{2,2}(\Omega)]^N} \\ &\leqq c\|g\|_{[W^{-1/2,2}(\partial\Omega)]^N} \|u_0\|_{[L^2(\Omega)]^N}, \end{aligned}$$

that is,

$$\|S_0(t)g\|_{[L^2(\Omega)]^N} \leqq c\|g\|_{[W^{-1/2,2}(\partial\Omega)]^N} \quad \forall g \in [W^{1/2,2}(\partial\Omega)]^N. \tag{2.58}$$

We now interpolate between (2.58) and (2.52), using Theorems 7.7 and 9.4 of [LM]: the proof of such theorems requires $\partial\Omega \in C^\infty$, but it can be readily adapted to our case. The result of interpolation is the estimate

$$\|S_0(t)g\|_{[W^{1/2,2}(\partial\Omega)]^N} \leqq c\|g\|_{[L^2(\partial\Omega)]^N} \quad \forall g \in [W^{1/2,2}(\partial\Omega)]^N, \tag{2.59}$$

which shows that the linear operator $S_0(t)$ may be boundedly extended to an operator $G_0(t) \in \mathscr{L}([L^2(\partial\Omega)]^N, [W^{1/2,2}(\Omega)]^N)$ defined by (compare with (2.47))

$$(2.60)\qquad \begin{aligned} &G_0(t): [L^2(\partial\Omega)]^N \to [W^{1/2,2}(\Omega)]^N,\\ &G_0(t)g := S_0(t)g \quad \forall g \in [W^{1/2,2}(\partial\Omega)]^N. \end{aligned}$$

We now turn to the case $r=1$. Fix $t \in [0, T]$, let $g \in [W^{1/2,2}(\partial\Omega)]^N$, and consider the problem corresponding to (2.9), (2.11):

$$(2.61)\qquad \begin{aligned} &\mathscr{A}(t, x, D)u_1 = 0 \quad \text{in } \Omega,\\ &\mathscr{B}_1(t, x, D)u_1 = g \quad \text{on } \partial\Omega, \end{aligned}$$

which, by [ADN], has a unique solution $u_1 := S_1(t)g \in [W^{2,2}(\Omega)]^N$, such that

$$(2.62)\qquad \|S_1(t)g\|_{[W^{2,2}(\Omega)]^N} \leqq c\|g\|_{[W^{1/2,2}(\partial\Omega)]^N} \quad \forall g \in [W^{1/2,2}(\partial\Omega)]^N.$$

Multiply by u_1 in $[L^2(\Omega)]^N$ in (2.61) and integrate by parts: the result is

$$\begin{aligned} \nu \int_\Omega |Du_1|^2\, dx + \int_\Omega |u_1|^2\, dx &\leqq \int_\Omega [(A_{sj}(t, x) \cdot D_j u_1 | D_s u_1)_{\mathbb{C}^N} + (u_1 | u_1)_{\mathbb{C}^N}]\, dx\\ &= \int_{\partial\Omega} (g | u_1)_{\mathbb{C}^N}\, d\sigma = |\langle g, u_1\rangle_{[W^{-1/2,2}(\partial\Omega)]^N, [W^{1/2,2}(\partial\Omega)]^N}|\\ &\leqq c\|g\|_{[W^{-1/2,2}(\partial\Omega)]^N} \|u_1\|_{[W^{1,2}(\Omega)]^N}, \end{aligned}$$

which implies

$$(2.63)\qquad \|S_1(t)g\|_{[W^{1,2}(\Omega)]^N} \leqq c\|g\|_{[W^{-1/2,2}(\partial\Omega)]^N} \quad \forall g \in [W^{1/2,2}(\partial\Omega)]^N.$$

Interpolation between (2.63) and (2.62) (see the remark after (2.58)) yields

$$(2.64)\qquad \|S_1(t)g\|_{[W^{3/2,2}(\Omega)]^N} \leqq c\|g\|_{[L^2(\partial\Omega)]^N} \quad \forall g \in [W^{1/2,2}(\partial\Omega)]^N,$$

i.e., S_1 may be boundedly extended to an operator $G_1(t) \in \mathscr{L}([L^2(\partial\Omega)]^N, [W^{3/2,2}(\Omega)]^N)$ defined by (compare with (2.48)):

$$(2.65)\qquad \begin{aligned} &G_1(t): [L^2(\partial\Omega)]^N \to [W^{3/2,2}(\Omega)]^N,\\ &G_1(t)g := S_1(t)g \quad \forall g \in [W^{1/2,2}(\partial\Omega)]^N. \end{aligned}$$

Now we recall that by Theorem 3.1 of [L1] we have for $r=0, 1$ (see (2.34)):

$$(2.66)\qquad D_{[-A_r(t)]^\vartheta} = D_{A_r(t)}(\vartheta, 2) \quad \forall \vartheta \in]0, 1[,$$

$$(2.67)\qquad D_{[-A_r(t)^*]^\vartheta} = D_{A_r(t)^*}(\vartheta, 2) \quad \forall \vartheta \in]0, 1[.$$

On the other hand, the real interpolation spaces $D_{A_r(t)}(\vartheta, 2)$ and $D_{A_r(t)^*}(\vartheta, 2)$ can be characterized in the following way:

$$(2.68)\qquad \begin{aligned} &D_{A_0(t)}(\vartheta, 2) = D_{A_0(t)^*}(\vartheta, 2)\\ &= \begin{cases} [W^{2\vartheta,2}(\Omega)]^N & \text{if } \vartheta \in]0, \tfrac14[,\\ \left\{u \in [W^{1/2,2}(\Omega)]^N : \displaystyle\int_\Omega d(x)^{-1}|u(x)|^2\, dx < \infty\right\} & \text{if } \vartheta = \tfrac14,\\ [W_0^{2\vartheta,2}(\Omega)]^N & \text{if } \vartheta \in]\tfrac14, 1[\setminus\{\tfrac12\},\\ [B_0^{1,2}(\Omega)]^N & \text{if } \vartheta = \tfrac12; \end{cases} \end{aligned}$$

(2.69)

$$D_{A_1(t)}(\vartheta, 2) = \begin{cases} [W^{2\vartheta,2}(\Omega)]^N & \text{if } \vartheta \in \left]0, \tfrac{3}{4}\right[\setminus\{\tfrac{1}{2}\}, \\ [B^{1,2}(\Omega)]^N & \text{if } \vartheta = \tfrac{1}{2}, \\ \left\{u \in [W^{3/2,2}(\Omega)]^N : \int_\Omega d(x)^{-1} \left| \sum_{sj=1}^n A_{sj}(t, x) \cdot D_j u(x) D_s d(x) \right|^2 dx < \infty \right\} & \\ & \text{if } \vartheta = \tfrac{3}{4}, \\ \{u \in [W^{2\vartheta,2}(\Omega)]^N : \mathscr{B}_1(t, \cdot, D)u = 0 \text{ on } \partial\Omega\} & \text{if } \vartheta \in \left]\tfrac{3}{4}, 1\right[; \end{cases}$$

(2.70)

$$D_{A_1(t)^*}(\vartheta, 2) = \begin{cases} [W^{2\vartheta,2}(\Omega)]^N & \text{if } \vartheta \in \left]0, \tfrac{3}{4}\right[\setminus\{\tfrac{1}{2}\}, \\ [B^{1,2}(\Omega)]^N & \text{if } \vartheta = \tfrac{1}{2}, \\ \left\{u \in [W^{3/2,2}(\Omega)]^N : \int_\Omega d(x)^{-1} \left| \sum_{sj=1}^n \overline{{}^t A_{sj}(t, x)} \cdot D_s u(x) D_j d(x) \right|^2 dx < \infty \right\} & \\ & \text{if } \vartheta = \tfrac{3}{4}, \\ \{u \in [W^{2\vartheta,2}(\Omega)]^N : \overline{\mathscr{B}_1(t, \cdot, D)}u = 0 \text{ on } \partial\Omega\} & \text{if } \vartheta \in \left]\tfrac{3}{4}, 1\right[. \end{cases}$$

Here $[B^{1,2}(\Omega)]^N$ is the Besov–Nikol'skij space. A proof of the results (2.68)–(2.70) is in Theorem 7.5 of [Gr] (see also [Tr, Thm. 4.3.3]) in the case $N = 1$ and $\partial\Omega \in C^\infty$, but the same argument works in our situation.

The above results (namely, (2.66)–(2.69) together with (2.60), (2.65)) show that

$$(2.71) \qquad \begin{aligned} G_0(t) &\in \mathscr{L}([L^2(\partial\Omega)]^N, D_{[-A_0(t)]^\vartheta}) \quad \forall \vartheta \in \left]0, \tfrac{1}{4}\right[, \\ G_1(t) &\in \mathscr{L}([L^2(\partial\Omega)]^N, D_{[-A_1(t)]^\vartheta}) \quad \forall \vartheta \in \left]0, \tfrac{3}{4}\right[; \end{aligned}$$

the norms of $G_0(t)$, $G_1(t)$ are bounded independently of $t \in [0, T]$ in view of (2.59), (2.64).

On the other hand, if we set

$$F_n(t) := [-A_r(t)]^\vartheta \exp\left(\frac{1}{n} A_r(t)\right) G_r(t),$$

we have $F_n \in C([0, T], \mathscr{L}([L^2(\partial\Omega)]^N, [L^2(\Omega)]^N))$ by Lemma 2.7(ii); in addition, choosing $\rho \in \left]0, \alpha_r - \vartheta\right[$ (with $\alpha_0 = \frac{1}{4}$, $\alpha_1 = \frac{3}{4}$) we see that

$$\begin{aligned} &\|F_n(t) - [-A_r(t)]^\vartheta G_r(t)\|_{\mathscr{L}([L^2(\partial\Omega)]^N, [L^2(\Omega)]^N)} \\ &\quad \leqq \left\| \int_0^{1/n} [-A_r(t)]^{1-\rho} \exp(\xi A_r(t))\, d\xi \right\|_{\mathscr{L}([L^2(\Omega)])^N} \\ &\quad\quad \cdot \|[-A_r(t)]^{\vartheta+\rho} G_r(t)\|_{\mathscr{L}([L^2(\partial\Omega)]^N, [L^2(\Omega)]^N)} \leqq \frac{c}{n^\rho}, \end{aligned}$$

so that $F_n(t) \to [-A_r(t)]^\vartheta G_r(t)$ in $\mathscr{L}([L^2(\partial\Omega)]^N, [L^2(\Omega)]^N)$ as $n \to \infty$, uniformly with respect to t; thus $[-A_r(\cdot)]^\vartheta G_r(\cdot)$ is a continuous function. This shows that

$$(2.72) \qquad [-A_0(\cdot)]^\vartheta G_0(\cdot) \in C([0, T], \mathscr{L}([L^2(\partial\Omega)]^N, [L^2(\Omega)]^N)) \quad \forall \vartheta \in \left]0, \tfrac{1}{4}\right[,$$

$$(2.73) \qquad [-A_1(\cdot)]^\vartheta G_1(\cdot) \in C([0, T], \mathscr{L}([L^2(\partial\Omega)]^N, [L^2(\Omega)]^N)) \quad \forall \vartheta \in \left]0, \tfrac{3}{4}\right[,$$

and, in particular, the proof is complete. □

We are ready to write a representation formula for (regular) solutions of problems (2.3), (2.4), which depends just on low-order norms, and hence can be extended to the case of less smooth data. This construction will be performed in the next section.

2.5. The representation formula. Consider again problems (2.3), (2.4) with smooth data: our representation formula for their solution is provided by the following proposition.

PROPOSITION 2.13. *Assume* (2.1), (2.2), *let* $y_0 \in [W^{2-r,2}(\Omega)]^N$ *and* $u \in C^{\alpha}([0, T], [W^{2-r,2}(\Omega)]^N) \cap C^{\alpha+1-r/2}([0, T], [L^2(\Omega)]^N)$ $(r=0$ *or* $r=1)$, *and suppose moreover that the compatibility conditions* (2.7) *or* (2.5) *hold. Then the solution of problem* (2.3), *or* (2.4), *is given by*

$$y(t, \cdot) = U_r(t, 0)y_0 + \int_0^t [[-A_r(s)^*]^{1-\vartheta} U_r(t, s)^*]^*[-A_r(s)]^{\vartheta} G_r(s)u(s, \cdot)\, ds, \tag{2.74}$$

$$t \in [0, T] \quad (\vartheta \in]0, \alpha_r[).$$

Proof. By Proposition 2.2 or 2.1 we know that problems (2.3) or (2.4) have a unique solution

$$y \in C^0([0, T], [W^{2,2}(\Omega)]^N) \cap C^1([0, T], [L^2(\Omega)]^N).$$

Consider the function $y - G_r(t)u$: by (2.47), (2.48), (2.53), and (2.62) we get (see (2.9))

$$\begin{aligned} &y(t, \cdot) - G_r(t)u(t, \cdot) \in D_{A_r(t)}, \\ &A_r(t)[y(t, \cdot) - G_r(t)u(t, \cdot)] = \mathscr{A}(t, \cdot, D)y(t, \cdot). \end{aligned} \tag{2.75}$$

Next, denoting by $U_r(t, s)$ the evolution operator associated to $\{A_r(t)\}_{t\in[0,T]}$, we have by Corollary 2.10

$$\|[-A_r(s)^*]^{\gamma} U_r(t, s)^*\|_{\mathscr{L}([L^2(\Omega)]^N)} \leqq M_{\gamma}(t-s)^{-\gamma} \quad \forall \gamma \in]0, 1[, \quad \forall 0 \leqq s < t \leqq T,$$

and consequently

$$\|[[-A_r(s)^*]^{\gamma} U_r(t, s)^*]^*\|_{\mathscr{L}([L^2(\Omega)]^N)} \leqq M_{\gamma}(t-s)^{-\gamma} \quad \forall \gamma \in]0, 1[, \quad \forall 0 \leqq s < t \leqq T. \tag{2.76}$$

Now fix $t \in [0, T]$, let $z \in D_{A_r(t)^*}$, and define

$$h(s) := (y(s, \cdot) \,|\, U_r(t, s)^* z)_{[L^2(\Omega)]^N}, \qquad s \in [0, t[. \tag{2.77}$$

By Proposition 2.9 and (2.75) we may compute

$$\begin{aligned} h'(s) &= (D_s y(s, \cdot) \,|\, U_r(t, s)^* z) - (y(s, \cdot) - G_r(s)u(s, \cdot) \,|\, A_r(s)^* U_r(t, s)^* z) \\ &\quad - (G_r(s)u(s, \cdot) \,|\, A_r(s)^* U_r(t, s)^* z) \\ &= (A(s, \cdot, D)y(s, \cdot) \,|\, U_r(t, s)^* z) \\ &\quad - (A_r(s)[y(s, \cdot) - G_r(s)u(s, \cdot)] \,|\, U_r(t, s)^* z) - (G_r(s)u(s, \cdot) \,|\, A_r(s)^* U_r(t, s)^* z) \\ &= -(G_r(s)u(s, \cdot) \,|\, A_r(s)^* U_r(t, s)^* z). \end{aligned}$$

On the other hand, by (2.72), (2.73) we may write for $\vartheta \in]0, \alpha_r[$ (with $\alpha_0 = \frac{1}{4}$, $\alpha_1 = \frac{3}{4}$)

$$\begin{aligned} h'(s) &= ([-A_r(s)]^{\vartheta} G_r(s)u(s, \cdot) \,|\, [-A_r(s)^*]^{1-\vartheta} U_r(t, s)^* z) \\ &= ([[-A_r(s)^*]^{1-\vartheta} U_r(t, s)^*]^*[-A_r(s)]^{\vartheta} G_r(s)u(s, \cdot) \,|\, z), \qquad s \in]0, t[, \end{aligned}$$

and $h' \in L^2(0, t)$. Hence by integrating in $]0, t[$ we get

$$\begin{aligned} &(y(t, \cdot) \,|\, z)_{[L^2(\Omega)]^N} - (U_r(t, 0)y_0(\cdot) \,|\, z)_{[L^2(\Omega)]^N} \\ &\quad = \left(\int_0^t [[-A_r(s)^*]^{1-\vartheta} U_r(t, s)^*]^*[-A_r(s)]^{\vartheta} G_r(s)u(s, \cdot)\, ds \,\Big|\, z \right)_{[L^2(\Omega)]^N}, \end{aligned}$$

and finally by density we deduce (2.74). □

Remark 2.14. (i) The representation formula (2.74) makes sense for any $y_0 \in [L^2(\Omega)]^N$ and $u \in [L^2(]0, T[\times \partial\Omega)]^N$, since by Proposition 2.2(i), (2.76), and (2.72), (2.73) we have

$$(2.78) \qquad \|y(t,\cdot)\|_{[L^2(\Omega)]^N} \leqq c\left\{\|y_0\|_{[L^2(\Omega)]^N} + \int_0^t (t-s)^{\vartheta-1}\|u(s,\cdot)\|_{[L^2(\partial\Omega)]^N}\,ds\right\},$$

which implies

$$(2.79) \qquad \|y\|^2_{[L^2(]0,T[\times\Omega)]^N} \leqq c\left\{T\|y_0\|^2_{[L^2(\Omega)]^N} + \frac{T^{2\vartheta}}{\vartheta^2}\|u\|^2_{[L^2(]0,T[\times\partial\Omega)]^N}\right\}.$$

(ii) We may rewrite formula (2.74) in a shorter, although improper, form, namely

$$(2.80) \qquad y(t) = U_r(t,0)y_0 - \int_0^t U_r(t,s)A_r(s)G_r(s)u(s)\,ds, \qquad t \in [0, T],$$

where the integrand is to be understood as in (2.74). In the foregoing section we will study an abstract version of (2.80) (see (3.1) below) within the context of control theory.

3. The L-Q-R problem over finite-time horizon.

3.1. State problem and cost functional. This section concerns the classical linear-quadratic regulator (L-Q-R) problem, over finite horizon $[0, T]$, for a class of abstract evolution equations corresponding to nonautonomous parabolic systems with boundary control. As we have shown in § 2, an equation of the form

$$(3.1) \qquad y(t) = U(t,0)y_0 - \int_0^t U(t,s)A(s)G(s)u(s)\,ds, \qquad t \in [0, T],$$

is appropriate to cover a wide class of concrete problems. In § 2 we derived in two concrete examples equation (2.80), which is an equation of the form (3.1), under hypotheses (2.1) and (2.2) (or, from the abstract point of view, (2.29) and (2.42)). Such assumptions will not be directly needed in most part of the next results on control problems; thus, in order to identify those properties which are really relevant from the control point of view, and to point out both analogies and novelties of the nonautonomous case with respect to the autonomous one (treated, e.g., in [B1], [La], [LT1], [LT2], [F1], [F2]), we will hereafter impose explicitly only assumptions (3.2)-(3.5) listed below.

Let H, U two separable (for simplicity) complex Hilbert spaces. In (3.1) we shall take $y_0 \in H$ and $u \in L^2(0, T; U)$. Here is our list of hypotheses:

(3.2) $\{A(t)\}_{t\in[0,T]}$ is a family of closed linear operators in H with (dense) domains $D_{A(t)}$, such that $A(t)$ generates an analytic semigroup in H and $0 \in \rho(A(t))$.

(3.3) $\{U(t,s)\}_{0\leqq s\leqq t\leqq T}$ is the (strongly continuous) evolution operator in H associated to $\{A(t)\}_{t\in[0,T]}$; in particular,

$$\|U(t,s)\|_{\mathscr{L}(H)} \leqq M_0, \text{ for all } (t,s) \in \bar{\Delta}, \text{ where } \Delta := \{(t,s) \in [0,T]^2 \colon s < t\}.$$

(3.4) The operator-valued function $(t,s) \to U(t,s)^*$ belongs to $C_s([0,T], \mathscr{L}(H))$; moreover, for each $\eta \in [0,1]$ and $(t,s) \in \Delta$, $U(t,s)^* \in \mathscr{L}(H, D_{[-A(s)^*]^\eta})$, the map $(t,s) \to [-A(s)^*]^\eta U(t,s)^*$ is strongly measurable and satisfies

$$\|[-A(s)^*]^\eta U(t,s)^*[-A(t)^*]^{-\mu}\|_{\mathscr{L}(H)} \leqq M_{\eta\mu}[(t-s)^{\mu-\eta}+1]$$

$$\forall (t,s) \in \Delta, \quad \forall \eta, \mu \in [0,1].$$

(3.5) $\{G(t)\}_{t\in[0,T]}$ is a family of operators in $\mathscr{L}(U, H)$ such that there exists $\alpha \in]0, 1]$ with the following properties: $G(t) \in \mathscr{L}(U, D_{[-A(t)^*]^\alpha})$ for each $t \in [0, T]$ and the map $t \to [-A(t)]^\alpha G(t)$ belongs to $L^\infty(0, T; \mathscr{L}(U, H))$.

Remark 3.1. (i) The above assumptions can be relaxed in various directions, with minor consequences on the subsequent results. So, for instance, (3.4) is needed only for $\eta = 1-\alpha$: in this case we would obtain slightly weaker regularity results for the Riccati equation. However, the applications discussed in § 2 do not motivate a further level of generality.

(ii) Condition (3.4) with $\mu > 0$ is not necessary to give sense to equation (3.1): just a much weaker version of it is needed in order to define the L-Q-R problem (3.10) below. However it will be used in (more or less) this generality as a technical tool in the study of the Riccati equation. Except for (3.4) with $\mu > 0$, all the other assumptions are the natural (and minimal, in a sense) ones in order to give a meaning to equation (3.1) and problem (3.10).

(iii) In the examples of § 2, we have under assumptions (2.1), (2.2):

$H = [L^2(\Omega)]^N$, $U = [L^2(\partial\Omega)]^N$;

$\{A(t)\}$, defined by (2.12) or (2.13), fulfills (3.2) by Proposition 2.4;

The existence of $\{U(t, s)\}$ with the properties (3.3) is guaranteed by Proposition 2.8(i);

Conditions (3.4) for $\{U(t, s)^*\}$ are proved in Corollary 2.10;

$\{G(t)\}$, defined by (2.47) or (2.48), satisfies (3.5) in view of Proposition 2.12.

As at the end of § 2, we agree that the formal notation $U(t, s)A(s)G(s)$ stands for $[[-A(s)^*]^{1-\alpha}U(t, s)^*]^*[-A(s)]^\alpha G(s)$, which is well defined as an element of $\mathscr{L}(U, H)$ for each $(t, s) \in \Delta$, by (3.4)-(3.5). More precisely we have Lemma 3.2.

LEMMA 3.2. *The operator-valued function*

$$U(t, s)A(s)G(s) := [[-A(s)^*]^{1-\alpha}U(t, s)^*]^*[-A(s)]^\alpha G(s), \qquad 0 \leqq s < t \leqq T, \tag{3.6}$$

is strongly measurable with respect to $s \in [0, t[$ *for each fixed* $t \in]0, T]$, *and strongly continuous with respect to* $t \in]s, T]$ *for each fixed* $s \in [0, T[$. *Moreover,*

$$\|U(t, s)A(s)G(s)\|_{\mathscr{L}(U,H)} \leqq c(t-s)^{\alpha-1} \quad \forall (t, s) \in \Delta. \tag{3.7}$$

Proof. The first assertion follows directly by (3.4), (3.5). Concerning the second one, let $s \in [0, T[$ and $t_0 \in]s, T]$ be fixed: it is easy to verify that if $t \in](s+t_0)/2, T]$ we have

$$U(t, s)A(s)G(s) = U(t, (s+t_0)/2)[U((s+t_0)/2, s)A(s)G(s)];$$

but $t \to U(t, (s+t_0)/2)$ is strongly continuous, whereas the bounded operator $U((s+t_0)/2, s)A(s)G(s)$ does not depend on t, so that $U(t, s)A(s)G(s)$ is strongly continuous at $t = t_0$. Finally, the estimate (3.7) follows by (3.4) and (3.5). □

The next lemma gives a precise interpretation of the function (3.1).

LEMMA 3.3. (i) *If* $u \in L^2(0, T; U)$, *then* (3.1) *defines a function* $y \in L^2(0, T; H)$ *and*

$$\|y\|_{L^2(0,T;H)} \leqq c\{\|y_0\|_H + \|u\|_{L^2(0,T;U)}\}. \tag{3.8}$$

(ii) *If* $u \in L^p(0, T; U)$ *for some* $p > 1/\alpha$, *then* $y \in C([0, T], H)$ *and*

$$\|y\|_{C([0,T],H)} \leqq c\{\|y_0\|_H + \|u\|_{L^p(0,T;U)}\}. \tag{3.9}$$

Proof. Part (i) is an easy consequence of (3.3), (3.7) and Young's inequality.

(ii) If $p>1/\alpha$, by (3.7) we have for $0\leqq r<t\leqq T$

$$\left\|\int_r^t U(t,s)A(s)G(s)u(s)\,ds\right\|_H \leqq \left[\int_r^t c(t-s)^{-(1-\alpha)p/(p-1)}\,ds\right]^{(p-1)/p}\|u\|_{L^p(r,T;U)}$$
$$\leqq c\frac{p-1}{\alpha p-1}(t-r)^{\alpha-1/p}\|u\|_{L^p(0,T;U)},$$

which, together with (3.30), implies in particular (3.9). Moreover, if $t_0\in\,]0,T]$ and $\varepsilon>0$, we have for small $\delta>0$

$$\left\|\int_{t_0-\delta}^t U(t,s)A(s)G(s)u(s)\,ds\right\|_H \leqq \varepsilon \quad \forall t\in[t_0-\delta, t_0+\delta].$$

Therefore by (3.3) we get for $|t-t_0|\leqq\delta$:

$$\begin{aligned}
&\|y(t)-y(t_0)\|_H\\
&\quad\leqq \|U(t,0)y_0-U(t_0,0)y_0\|_H\\
&\qquad+\left\|[U(t,t_0-\delta)-U(t_0,t_0-\delta)]\int_0^{t_0-\delta}U(t_0-\delta,s)A(s)G(s)u(s)\,ds\right\|_H\\
&\qquad+\left\|\int_{t_0-\delta}^t U(t,s)A(s)G(s)u(s)\,ds\right\|_H+\left\|\int_{t_0-\delta}^{t_0} U(t_0,s)A(s)G(s)u(s)\,ds\right\|_H\\
&\quad\leqq \|U(t,0)y_0-U(t_0,0)y_0\|_H+(2M_0+2)\varepsilon,
\end{aligned}$$

and the result follows by the strong continuity of $t\to U(t,0)$. The case $t_0=0$ is even simpler. □

We can now define the optimal control problem which is the object of our study in this section. We shall consider the following L-Q-R problem:

(3.10) Minimize

$$J(u):=\int_0^T[(M(t)y(t)|y(t))_H+(N(t)u(t)|u(t))_U]\,dt+(P_Ty(T)|y(T))_H$$

over all controls $u\in L^2(0,T;U)$ subject to the state equation (3.1).

Here we assume:

(3.11) $M(t)\in\Sigma^+(H)$, for all $t\in[0,T]$ and $M\in L^\infty(0,T;\mathscr{L}(H))$;

(3.12) $N(t)\in\Sigma^+(U)$ with $N(t)\geqq\nu>0$, for all $t\in[0,T]$ and $N\in C_s([0,T],\mathscr{L}(U))$;

(3.13) $P_T\in\Sigma^+(H)$, and there exists $\beta\in\,](\frac{1}{2}-\alpha),\frac{1}{2}]\cap[0,\frac{1}{2}]$ such that $P_T\in\mathscr{L}(H,D_{[-A(T)^*]^{2\beta}})$.

Remark 3.4. Due to Lemma 3.3 of [F1], assumption (3.13) implies that the operator $[-A(T)^*]^{\beta-\varepsilon}P_T[-A(T)]^{\beta-\varepsilon}(\varepsilon\in\,]0,\beta])$ can be extended to an operator $L_\varepsilon\in\mathscr{L}(H)$.

Note that y, given by (3.1), is not continuous in general, but only in $L^2(0,T;H)$: hence the term $(P_Ty(T)|y(T))_H$ is not well defined a priori for all controls $u\in L^2(0,T;U)$, but only for controls in a dense subspace of $L^2(0,T;U)$, by Lemma 3.3(ii). However, the regularity property (3.13), along with (3.4), yields Lemma 3.5.

LEMMA 3.5. *The mapping $u\to(P_Ty(T)|y(T))_H$, defined (for instance) from $C([0,T],U)$ into $\mathbb{R}$, is locally uniformly continuous with respect to the topology of $L^2(0,T;U)$, and hence it can be extended to $L^2(0,T;U)$.*

Proof. Let L_ε be the bounded extension to H of the operator $[-A(T)^*]^{\beta-\varepsilon}P_T[-A(T)]^{\beta-\varepsilon}$ (see Remark 3.4). If $u\in C([0,T],U)$, choosing $\varepsilon\in\,]0,\beta-(\frac{1}{2}-\alpha)[$ we have by (3.1), (3.3), (3.4), and (3.5)

$$\begin{aligned}
&(P_Ty(T)\,|\,y(T))_H\\
&\quad=\|L_\varepsilon^{1/2}[-A(T)]^{-\beta+\varepsilon}y(T)\|_H^2\\
&\quad\leqq\|L_\varepsilon^{1/2}\|^2_{\mathscr{L}(H)}\Bigg\{\|[-A(T)]^{-\beta+\varepsilon}U(T,0)y_0\|_H\\
&\qquad\qquad+\int_0^T\|[[-A(s)^*]^{1-\alpha}U(T,s)^*[-A(T)^*]^{-\beta+\varepsilon}]^*[-A(s)]^\alpha G(s)u(s)\|_H\,ds\Bigg\}^2\\
&\quad\leqq c\left\{\|y_0\|_H^2+\left[\int_0^T[1+(T-s)^{\beta+\alpha-\varepsilon-1}]\|u(s)\|_U\,ds\right]^2\right\}\\
&\quad\leqq c\{\|y_0\|_H^2+(T+T^{2\beta+2\alpha-2\varepsilon-1})\|u\|^2_{L^2(0,T;U)}\}.
\end{aligned}$$

Hence if u_1, $u_2\in C([0,T],U)$ and y_1, y_2 are the corresponding functions (3.1) with initial state y_0 we have

$$\begin{aligned}
&|(P_Ty_1(T)\,|\,y_1(T))_H-(P_Ty_2(T)\,|\,y_2(T))_H|\\
&\quad\leqq[\|L_\varepsilon^{1/2}[-A(T)]^{-\beta+\varepsilon}y_1(T)\|_H+\|L_\varepsilon^{1/2}[-A(T)]^{-\beta+\varepsilon}y_2(T)\|_H]\\
&\qquad\cdot\|L_\varepsilon^{1/2}[-A(T)]^{-\beta+\varepsilon}[y_1(T)-y_2(T)]\|_H\\
&\quad\leqq c\{\|y_0\|_H+\|u_1\|_{L^2(0,T;U)}+\|u_2\|_{L^2(0,T;U)}\}\|u_1-u_2\|^2_{L^2(0,T;U)}.\qquad\square
\end{aligned}$$

Remark 3.6. The initial state y_0 can be taken in a space larger than H without changing the main results of this and subsequent sections. More precisely, we need to fulfill two essential requirements, namely (1°) $y\in L^2(0,T;H)$, and (2°) $[-A(T)]^{-\beta}y(T)$ is well defined; in order to get them, it is sufficient that $[-A(0)]^{-\delta}y_0\in H$ for some $\delta\in\,]0,\frac{1}{2}[$, i.e., that y_0 belongs to the dual of $D_{[-A(0)^*]^\delta}$ with respect to H (indeed $[-A(0)]^{-\delta}$ can be extended to an isomorphism between the dual of $D_{[-A(0)^*]^\delta}$ and H). In this case the condition $y\in L^2(0,T;H)$ is satisfied because of (3.4) (with $\mu=0$), since we have

$$\begin{aligned}
(U(t,0)y_0\,|\,x)_H&=([-A(0)]^{-\delta}y_0\,|\,[-A(0)^*]^\delta U(t,0)^*x)_H\\
&\leqq\|[-A(0)]^{-\delta}y_0\|_H M_{\delta 0}t^{-\delta}\|x\|_H\quad\forall x\in H,
\end{aligned}$$

i.e., $\|U(t,0)y_0\|_H\leqq ct^{-\delta}$, $\delta\in\,]0,\frac{1}{2}[$; on the other hand, $[-A(T)]^{-\beta}y(T)$ is well defined (even if $\delta\in[\frac{1}{2},1[$) by (3.4), since

$$[-A(T)]^{-\beta}U(T,0)y_0=[[-A(0)^*]^\delta U(T,0)^*[-A(T)^*]^{-\beta}]^*[-A(0)]^{-\delta}y_0.$$

3.2. The Riccati equation. The main step in the solution of problem (3.10) is the direct study of the associated Riccati equation, which takes the form

$$\begin{aligned}
P(t)=U(T,t)^*P_TU(T,t)+\int_t^T U(s,t)^*&\\
\cdot[M(s)-P(s)A(s)G(s)N(s)^{-1}G(s)^*A(s)^*P(s)]U(s,t)\,ds.&
\end{aligned}\tag{3.14}$$

The nonlinear term in (3.14) is not well defined in the present form. For this reason we consider the following version of (3.14):

$$P(t) = U(T, t)^* P_T U(T, t) + \int_t^T U(s, t)^* \cdot [M(s) - [[-A(s)^*]^{1-\alpha} P(s)]^* K(s) [-A(s)^*]^{1-\alpha} P(s)] U(s, t)\, ds, \tag{3.15}$$

where

$$K(s) := [-A(s)]^{\alpha} G(s) N(s)^{-1} [[-A(s)]^{\alpha} G(s)]^*. \tag{3.16}$$

By (3.5) and (3.12) we have

$$K(\cdot) \in L^{\infty}(0, T; \Sigma^+(H)). \tag{3.17}$$

Note that the integration in (3.15) is performed in the strong sense.

PROPOSITION 3.7 (local solution). *There exist an interval $[T_0, T]$ and a unique function $P \in C_s([T_0, T], \Sigma(H))$ such that*:

(i) $[-A(\cdot)^*]^{1-\alpha} P(\cdot)$ *is well defined and strongly measurable from* $[T_0, T]$ *into* $\mathscr{L}(H)$,

(ii) $\|[-A(t)^*]^{1-\alpha} P(t)\|_{\mathscr{L}(H)} \leqq c(T-t)^{-(1-\alpha-2\beta)\vee 0}$, *for all* $t \in [T_0, T[$,

(iii) $P(\cdot)$ *solves* (3.15) *in* $[T_0, T]$.

Proof. For any $T_0 \in [0, T[$ denote by $B_\gamma(T_0, T)$ the Banach space of all strongly measurable functions $Q: [T_0, T[\to \mathscr{L}(H)$ such that

$$\|Q\|_{B_\gamma(T_0,T)} := \sup_{T_0 \leqq t \leqq T} (T-t)^{\gamma} \|Q(t)\|_{\mathscr{L}(H)} < \infty,$$

where $\gamma := (1-\alpha-2\beta) \vee 0$. For $Q \in B_\gamma(T_0, T)$, define

$$\Gamma_{T_0}(Q)(t) := [-A(t)^*]^{1-\alpha} U(T, t)^* P_T U(T, t) + \int_t^T [-A(t)^*]^{1-\alpha} U(s, t)^* [M(s) - Q(s)^* K(s) Q(s)] U(s, t)\, ds, \qquad t \in [T_0, T].$$

Let us show that Γ_{T_0} maps $B_\gamma(T_0, T)$ into itself. By (3.4), (3.13), (3.3), (3.11), and (3.17) we get

$$\begin{aligned} &\|\Gamma_{T_0}(Q)(t)\|_{\mathscr{L}(H)} \\ &\quad \leqq M_{1-\alpha,2\beta}[1 + (T-t)^{2\beta+\alpha-1}] \|[-A(T)^*]^{2\beta} P_T\|_{\mathscr{L}(H)} \|U(T, t)\|_{\mathscr{L}(H)} \\ &\qquad + M_{1-\alpha,0} \int_t^T (s-t)^{\alpha-1} [\|M(s)\|_{\mathscr{L}(H)} \\ &\qquad + \|K(s)\|_{\mathscr{L}(H)} (T-s)^{-2\gamma} \|Q\|^2_{B_\gamma(T_0,T)}] \|U(s, t)\|_{\mathscr{L}(H)}\, ds \\ &\quad \leqq c[(T-t)^{-\gamma} + (T-t)^{\alpha} + (T-t)^{\alpha-2\gamma} \|Q\|^2_{B_\gamma(T_0,T)}] \quad \forall t \in [T_0, T[; \end{aligned}$$

this shows that $\Gamma_{T_0}(Q) \in B_\gamma(T_0, T)$ and

$$\|\Gamma_{T_0}(Q)\|_{B_\gamma(T_0,T)} \leqq c_1 + c_2 (T - T_0)^{\alpha-\gamma} \|Q\|^2_{B_\gamma(T_0,T)}. \tag{3.18}$$

Next, we show that Γ_{T_0} is a contradiction in the ball

$$B_\gamma(T_0, T; \rho) := \{Q \in B_\gamma(T_0, T) : \|Q\|_{B_\gamma(T_0,T)} \leqq \rho\}$$

for a suitable $\rho > 0$. Indeed if $Q_1, Q_2 \in B_\gamma(T_0, T; \rho)$ we have as before:

$$\begin{aligned}\|\Gamma_{T_0}(Q_1)(t) - \Gamma_{T_0}(Q_2)(t)\|_{\mathscr{L}(H)} &\\ \leqq c \int_t^T (s-t)^{\alpha-1}[\|Q_1(s)\|_{\mathscr{L}(H)} + \|Q_2(s)\|_{\mathscr{L}(H)}]\|Q_1(s) - Q_2(s)\|_{\mathscr{L}(H)}\, ds &\\ \leqq c(T-t)^{\alpha-2\gamma}\rho\|Q_1 - Q_2\|_{B_\gamma(T_0,T)}, \quad \forall t \in [T_0, T[, &\end{aligned}$$

which implies

$$\|\Gamma_{T_0}(Q_1) - \Gamma_{T_0}(Q_2)\|_{B_\gamma(T_0,T)} \leqq c\rho(T-t)^{\alpha-\gamma}\|Q_1 - Q_2\|_{B_\gamma(T_0,T)}. \tag{3.19}$$

By (3.18) and (3.19) we see that it is possible to choose a (large) $\rho > 0$ and a $T_0 \in [0, T[$ (close to T), such that Γ_{T_0} maps $B_\gamma(T_0, T; \rho)$ into itself and is a contraction in $B_\gamma(T_0, T; \rho)$. Thus we get a unique solution of the equation

$$Q = \Gamma_{T_0}(Q) \quad \text{in } [T_0, T[.$$

Hence $P := [-A(\cdot)^*]^{\alpha-1}Q(\cdot)$ is the unique operator-valued function that satisfies (i)-(iii). The property $P \in C_s([T_0, T], \mathscr{L}(H))$ follows by (3.15), whereas the property $P(t) \in \Sigma(H)$ is a consequence of the fact that $P(\cdot)^*$ is also in $C_s([T_0, T], \mathscr{L}(H))$ and satisfies (i)-(iii), so that $P(t)^* \equiv P(t)$ in $[T_0, T]$. □

Remark 3.8. Since $[-A(T_0)]^{1-\alpha}P(T_0) \in \mathscr{L}(H)$, for each $\beta \in]\frac{1}{2} - \alpha, (1-\alpha)/2[\cap [0, (1-\alpha)/2[$ the operator $[-A(T_0)^*]^\beta P(T_0)[-A(T_0)]^\beta$ is continuous with respect to the topology of H (see [F1, Lemma 3.3]).

The result of Proposition 3.7 justifies the following definition:

DEFINITION 3.9. Let J be an interval in $[0, T]$ such that $T \in J$. We say that P is a solution of (3.15) in J if:

(i) $P \in C_s(J, \Sigma(H))$, $[-A(\cdot)^*]^{1-\alpha}P(\cdot)$ is well defined and strongly measurable from J into $\mathscr{L}(H)$,

(ii) For each $\tau \in J \backslash \{T\}$ there exists a constant $c(\tau)$ such that

$$\|[-A(t)^*]^{1-\alpha}P(t)\|_{\mathscr{L}(H)} \leqq c(\tau)(T-t)^{-\gamma} \quad \forall t \in [\tau, T[,$$

where $\gamma = (1 - \alpha - 2\beta) \vee 0$,

(iii) $P(\cdot)$ satisfies (3.15) in J.

We must prove the existence and uniqueness of a global solution, i.e., a solution in $[0, T]$, of (3.15). The proof will be based on an a priori bound; to this purpose we introduce an evolution operator which will be related to the optimal trajectories of problem (3.10).

LEMMA 3.10. *Let P be a solution of* (3.15) *in J. Consider the integral equation*

$$\begin{aligned}\Phi(t,s)x = U(t,s)x + \int_s^t & [[-A(r)^*]^{1-\alpha}U(t,r)^*]^*K(r)\\ &\cdot [[-A(r)^*]^{1-\alpha}P(r)]^*\Phi(r,s)x\, dr, \qquad t \in J,\end{aligned} \tag{3.20}$$

where $x \in H$. Then there exists a unique operator-valued function $\Phi : \overline{\Delta_J} \to \mathscr{L}(H)$, with $\Delta_J := \{(t,s) \in J^2 : t > s\}$, such that $\Phi(t,s)x$ is a solution in $C([s,T],H)$ of (3.20) *for each $x \in H$ and $s \in J$. Moreover, Φ is a strongly continuous evolution operator.*

Proof. Fix $s \in J$. If $\Phi(\cdot, s)x \in C([s, T], H)$, then

$$[-A(\cdot)^*]^{1-\alpha}P(\cdot)\Phi(\cdot, s)x \in L^p(s, T; H)$$

for some $p > 1/\alpha$ (by Definition 3.9(ii)); thus (3.16), (3.5), (3.12), and Lemma 3.3(i) imply that the right-hand side of (3.20) is in $C([s, T], H)$. Therefore it is standard to apply the contraction principle to (3.20), in order to get existence of a unique solution of (3.20) in $C([s, T], H)$, denoted by $\Phi(\cdot, s)x$. The proof that Φ is a strongly continuous evolution operator is classical. □

Using the evolution operator Φ it is possible to rewrite the Riccati equation (3.15) in two alternative integral forms as follows.

LEMMA 3.11. *If P is a solution of* (3.15) *in J and $\bar{T} \in J$, then for each $t \in J \cap [0, \bar{T}]$*

$$P(t) = \Phi(\bar{T}, t)^* P(\bar{T}) \Phi(\bar{T}, t) + \int_t^{\bar{T}} \Phi(s, t)^* \cdot [M(s) + [[-A(s)^*]^{1-\alpha} P(s)]^* K(s) [-A(s)^*]^{1-\alpha} P(s)] \Phi(s, t)\, ds, \tag{3.21}$$

$$P(t) = U(\bar{T}, t)^* P(\bar{T}) \Phi(\bar{T}, t) + \int_t^{\bar{T}} U(\sigma, t)^* M(\sigma) \Phi(\sigma, t)\, d\sigma. \tag{3.22}$$

Proof. The proof is classical (see, e.g., [Gi], [LT1]). □

We are now able to prove the following a priori bound, which is the key point in showing global existence.

LEMMA 3.12. *There exists $c > 0$ with the following property: if P is a solution of* (3.15) *in some interval J, then*

$$\|[-A(t)^*]^{1-\alpha} P(t)\|_{\mathscr{L}(H)} \leqq c(T-t)^{-\gamma} \quad \forall t \in J \backslash \{T\}, \tag{3.23}$$

where $\gamma = (1 - \alpha - 2\beta) \vee 0$.

Proof. Of course (3.23) is obvious if $t \in J \cap [T_0, T[$, with T_0 given by Proposition 3.7. Thus we may confine ourselves to consider the interval $J \cap [0, T_0]$.

Our first step consists in showing that there exists $c > 0$, independent of P and J, such that

$$\|P(t)\|_{\mathscr{L}(H)} \leqq c \quad \forall t \in J. \tag{3.24}$$

Indeed, choose $\bar{T} = T$ in Lemma 3.11: by (3.21) we have

$$P(t) \geqq 0 \quad \forall t \in J; \tag{3.25}$$

moreover, by (3.15) we get

$$\begin{aligned}(P(t)x \mid x)_H &\leqq \|P_T^{1/2} U(T, t)x\|_H^2 + \int_t^T \|M(s)^{1/2} U(s, t)x\|_H^2\, ds \\ &\leqq c\|x\|_H^2 \quad \forall x \in H, \quad \forall t \in J,\end{aligned}$$

with c independent of P and J. Thus (3.24) follows by (3.25). Next, by (3.22) we deduce for $s, t \in J \cap [0, T_0]$, $s \leqq t$:

$$\begin{aligned}[-A(t)^*]^{1-\alpha} P(t) \Phi(t, s) &= [-A(t)^*]^{1-\alpha} U(T_0, t)^* P(T_0) \Phi(T_0, s) \\ &\quad + \int_t^{T_0} [-A(t)^*]^{1-\alpha} U(\sigma, t)^* M(\sigma) \Phi(\sigma, s)\, d\sigma \\ &=: I_1(t, s) + I_2(t, s),\end{aligned} \tag{3.26}$$

where T_0 is taken as in Proposition 3.7. By (3.21) and (3.24) we obtain a first estimate:

$$\int_s^{T_0} \|K(t)^{1/2} [-A(t)^*]^{1-\alpha} P(t) \Phi(t, s)x\|_H^2\, dt \leqq c\|x\|_H^2 \quad \forall x \in H, \quad \forall s \in J \cap [0, T_0], \tag{3.27}$$

with c independent of $s \in J \cap [0, T_0]$ and $x \in H$.

The proof now proceeds in the following manner. Starting from (3.27), we will apply a bootstrap process in order to get more and more summability for the function $[-A(\cdot)^*]^{1-\alpha}P(\cdot)\Phi(\cdot, s)x$ in the interval $J \cap [s, T_0]$, where $s \in J \cap [0, T_0]$. Our final goal is the estimate

$$(3.28)\quad \|[-A(t)^*]^{1-\alpha}P(t)\Phi(t, s)x\|_H \leqq c\|x\|_H \quad \forall x \in H, \quad \forall s, t \in J \cap [0, T_0], \quad s \leqq t,$$

with c independent of x, P, s, t, and J: choosing in (3.28) $s = t$, (3.23) will follow, thus completing the proof of Lemma 3.12.

The bootstrap procedure works as follows. Let $p \in [2, \infty[$ be given, and set

$$p_0 := \begin{cases} \dfrac{p}{1-\alpha p} & \text{if } p \in \left[2, \dfrac{1}{\alpha}\right], \\ +\infty & \text{if } p \geqq \dfrac{1}{\alpha}. \end{cases}$$

Clearly, if $\alpha \in [\frac{1}{2}, 1]$ we have $1/\alpha \leqq 2$ so that $p_0 = +\infty$ whatever be p. If otherwise $\alpha \in]0, \frac{1}{2}[$, then

$$(3.29)\qquad p_0 - p = \frac{\alpha p^2}{1-\alpha p} \geqq \frac{4\alpha}{1-2\alpha} > 0 \quad \forall p \in \left[2, \frac{1}{\alpha}\right].$$

Assuming the truth of the estimate (c independent of x, P, s, J)

$$(3.30)\quad \|[-A(\cdot)^*]^{1-\alpha}P(\cdot)\Phi(\cdot, s)x\|_{L^p(s,T_0;H)} \leqq c\|x\|_H \quad \forall x \in H, \quad \forall s \in J \cap [0, T_0],$$

we will prove the same estimate with p replaced by p_0. This argument starts with $p = 2$, in which case we assume (3.27) instead of (3.30), and stops after a finite number of iterations (by virtue of (3.29)), the final estimate being (3.28). Suppose that (3.30) holds for a certain $p \geqq 2$: by (3.26) it is enough to show that

$$(3.31)\qquad \|I_1(\cdot, s)\|_{L^{p_0}(s,T_0;H)} \leqq c\|x\|_H,$$

$$(3.32)\qquad \|I_2(\cdot, s)\|_{L^{p_0}(s,T_0;H)} \leqq c\|x\|_H.$$

Concerning (3.31), by (3.27), using (3.3), (3.4), and (3.17) we get

$$\int_s^{T_0} \|\Phi(t, s)x\|_H^{p_0}\, dt$$

$$\leqq c\left\{\|x\|_H^{p_0} + \int_s^{T_0}\left[\int_s^t (t-r)^{\alpha-1}\|[-A(r)^*]^{1-\alpha}P(r)\Phi(r, s)x\|_H\, dr\right]^{p_0} dt\right\}$$

and a Young-type estimate [HLP, Thm. 383], along with our assumption (3.30), yields

$$\int_s^{T_0} \|\Phi(t, s)x\|_H^{p_0}\, dt \leqq c\|x\|_H^{p_0},$$

with c independent of P and J (with obvious modifications if $p_0 = \infty$). The bound (3.31) now follows by (3.4) and (3.11), applying the simplest version of Young's inequality.

Let us verify (3.32). First we observe that for any $\eta \in [0, 1[$ and $\varepsilon \in]0, 1-\eta[$ the operator $[-A(T_0)^*]^{1-\eta-\varepsilon}P(T_0)[-A(T_0)]^\eta$ can be uniquely extended to a bounded linear operator in H. Indeed, by (3.15) and (3.13) we have for each $x \in D_{[-A(T_0)]^\eta}$

$$\begin{aligned}
&[-A(T_0)^*]^{1-\eta-\varepsilon} P(T_0)[-A(T_0)]^\eta x \\
&\quad = [[-A(T_0)^*]^{1-\eta-\varepsilon} U(T, T_0)^*[-A(T)^*]^{-2\beta}][[-A(T)^*]^{2\beta} P_T] \\
&\qquad \cdot [[-A(T_0)^*]^\eta U(T, T_0)^*]^* x + \int_{T_0}^{T} [-A(T_0)^*]^{1-\eta-\varepsilon} U(s, T_0)^* \\
&\qquad \cdot [M(s) - [[-A(s)^*]^{1-\alpha} P(s)]^* K(s)[-A(s)^*]^{1-\alpha} P(s)] \\
&\qquad \cdot [[-A(T_0)^*]^\eta U(s, T_0)^*]^* x \, ds,
\end{aligned}$$

and hence by (3.4), (3.13), (3.11), and Proposition 3.7

$$\begin{aligned}
&\|[-A(T_0)^*]^{1-\eta-\varepsilon} P(T_0)[-A(T_0)]^\eta x\|_H \\
&\quad \leqq c\left\{[1+(T-T_0)^{2\beta-1+\eta+\varepsilon}][1+(T-T_0)^{-\eta}] + \int_{T_0}^{T} (s-T_0)^{\varepsilon-1}(T-s)^{-2\gamma}\, ds\right\} \|x\|_H \\
&\quad \leqq c\|x\|_H
\end{aligned}$$

(with c independent of x, P, J). Moreover, if $\eta > 1/p - \alpha$ we have by (3.20), (3.3), (3.4), (3.17), and (3.30)

$$\begin{aligned}
\|[-A(T_0)]^{-\eta}\Phi(T_0, s)x\|_H &\leqq c\|x\|_H + c\left[\int_s^{T_0} (T_0 - r)^{(\alpha-1+\eta)p/(p-1)}\, dr\right]^{(p-1)/p} \\
&\qquad \cdot \|[-A(\cdot)]^{1-\alpha} P(\cdot)\Phi(\cdot, s)x\|_{L^p(0,T_0;H)} \leqq c\|x\|_H
\end{aligned}$$

(c independent of x, P, s, J). Therefore we can rewrite $I_1(t, s)$ as

$$\begin{aligned}
I_1(t, s) &= [[-A(t)^*]^{1-\alpha} U(T_0, t)^*[-A(T_0)^*]^{\eta+\varepsilon-1}] \\
&\qquad \cdot [[-A(T_0)^*]^{1-\eta-\varepsilon} P(T_0)[-A(T_0)]^\eta][[-A(T_0)]^{-\eta}\Phi(T_0, s)];
\end{aligned}$$

hence if we take $\eta \in]1/p - \alpha, 1/p[$ and $\varepsilon \in]0, 1/p[$ we see using (3.4) that

$$\begin{aligned}
&\int_s^{T_0} \|[-A(t)^*]^{1-\alpha} U(T_0, t)^*[-A(T_0)^*]^{\eta+\varepsilon-1} x\|_H^{p_0}\, dt \\
&\qquad \leqq c \int_s^{T_0} (T_0 - t)^{(\alpha-\eta-\varepsilon)p_0}\, dt \|x\|_H^{p_0} \leqq c\|x\|_H^{p_0},
\end{aligned}$$

since $(\eta + \varepsilon - \alpha)p_0 < 1$. Thus we immediately obtain (3.32). Hence we get (3.30) with p replaced by p_0; consequently (3.28) follows, and the proof of Lemma 3.12 is complete. □

We can now prove the main result of this section.

THEOREM 3.13 (global solution). *There exists a unique solution P of equation* (3.15) *in* $[0, T]$. *Moreover, it has the following properties*:

(i) $P(t) \geqq 0$ *for each* $t \in [0, T]$;

(ii) *P satisfies the integral Riccati equations* (3.21) *and* (3.22);

(iii) *P satisfies the bounds* (3.24) *in* $[0, T]$ *and* (3.23) *in* $[0, T[$;

(iv) *For each $\eta \in [0, 1[$, the linear operator $[-A(t)^*]^\eta P(t)$, $t \in [0, T[$, is well defined, strongly measurable in t, and equibounded on compact subsets of* $[0, T[$.

Proof. Let T_0 be given by Proposition 3.7. For each $T_1 \in [0, T_0[$, consider the Banach space $L^\infty(T_1, T_0; \mathscr{L}(H))$ and the balls

$$B(T_1, T_0; \rho) := \{Q \in L^\infty(T_1, T_0; \mathscr{L}(H)) : \|Q\|_{L^\infty(T_1, T_0; \mathscr{L}(H))} \leqq \rho\}, \qquad \rho > 0.$$

Define the mapping Γ_{T_1,T_0} on $L^\infty(T_1, T_0; \mathscr{L}(H))$ by

$$\begin{aligned}\Gamma_{T_1,T_0}(Q)(t) := {} & [[-A(t)^*]^{1-\alpha}U(T_0, t)^*[-A(T_0)^*]^{\alpha-1}][[-A(T_0)^*]^{1-\alpha}P(T_0)U(T_0, t)] \\ & + \int_t^{T_0} [-A(t)^*]^{1-\alpha}U(s, t)^*[M(s) - Q(s)^*K(s)Q(s)]U(s, t)\, ds, \\ & \qquad t \in [T_1, T_0[,\end{aligned}$$

$$\Gamma_{T_1,T_0}(Q)(T_0) := [-A(T_0)^*]^{1-\alpha}P(T_0).$$

As in the proof of Proposition 3.7, we have

$$\begin{aligned}\|\Gamma_{T_1,T_0}(Q)\|_{L^\infty(T_1,T_0;\mathscr{L}(H))} \leqq {} & c_1\|[-A(T_0)^*]^{1-\alpha}P(T_0)\|_{\mathscr{L}(H)} + c_2 \\ & + c_3(T_0 - T_1)^\alpha \|Q\|^2_{L^\infty(T_1,T_0;\mathscr{L}(H))}, \\ & \forall Q \in L^\infty(T_1, T_0; \mathscr{L}(H)),\end{aligned} \tag{3.33}$$

$$\begin{aligned}&\|\Gamma_{T_1,T_0}(Q_1) - \Gamma_{T_1,T_0}(Q_2)\|_{L^\infty(T_1,T_0;\mathscr{L}(H))} \\ &\leqq c_4\rho(T_0 - T_1)^\alpha \|Q_1 - Q_2\|_{L^\infty(T_1,T_0;\mathscr{L}(H))} \quad \forall Q_1, Q_2 \in L^\infty(T_1, T_0; \mathscr{L}(H));\end{aligned} \tag{3.34}$$

here $c_1, \cdots, c_4$ are constants independent of T_1, T_0. Using the a priori bound (3.23), by (3.33) and (3.34) we see that we can select $\rho > 0$ and $T_1 \in]0, T_0[$ such that:

(a) $T_0 - T_1$ and ρ are independent of T_0;

(b) Γ_{T_1,T_0} is a contraction which maps $B(T_1, T_0; \rho)$ into itself. Thus there exists a unique solution Q of the equation

$$Q = \Gamma_{T_1,T_0}(Q)$$

in the space $L^\infty(T_1, T_0; \mathscr{L}(H))$, and this procedure can be repeated in the interval $[T_1 - (T_0 - T_1), T_1]$, and so on, with constant step. As in the proof of Proposition 3.7, we conclude that there exists a unique solution P of (3.15) in $[0, T]$.

Finally, property (i) follows by (3.25), and similarly properties (ii) and (iii) are proved in Lemmas 3.11 and 3.12. As to (iv), it is sufficient to use (3.4), (3.11) and the last assertion of Lemma 3.10 in equation (3.22) with $\bar{T} = T$. □

3.3. Synthesis. The results of the preceding section lead to the following theorem.

THEOREM 3.14. *Let $y_0 \in H$ be given. Then*:

(i) *There exists a unique optimal control $\hat{u}_0 \in L^2(0, T; U)$ for problem* (3.10);

(ii) *Denoting by $P(\cdot)$ the solution of the Riccati equation* (3.15), *we have*

$$J(\hat{u}_0) = (P(0)y_0|y_0)_H; \tag{3.35}$$

(iii) *If $\hat{y}_0 \in L^2(0, T; H)$ is the optimal trajectory, i.e., the solution of the state equation* (3.1) *corresponding to $\hat{u}_0(\cdot)$, we have the feedback formula for $\hat{u}_0(\cdot)$*:

$$\hat{U}_0(t) = N(t)^{-1}G(t)^*A(t)^*P(t)\hat{Y}_0(t), \qquad t \in [0, T[; \tag{3.36}$$

(iv) *The optimal trajectory $\hat{y}_0(\cdot)$ is expressed by*

$$\hat{y}_0(t) = \Phi(t, 0)y_0, \tag{3.37}$$

where $\Phi(t, s)$ is defined by the integral equation (3.20) *with $J = [0, T]$*;

(v) *The optimal pair $(\hat{u}_0, \hat{y}_0)$ is characterized by the following optimality system*:

$$\begin{aligned}&\hat{Y}_0(t) = U(t, 0)Y_0 - \int_0^t U(t, s)A(s)G(s)\hat{U}_0(s)\, ds, \\ &\hat{u}_0(t) = N(t)^{-1}G(t)^*A(t)^*p(t), \qquad t \in [0, T[, \\ &p(t) = U(T, t)^*P_T\hat{y}_0(T) + \int_t^T U(s, t)^*M(s)\hat{y}_0(s)\, ds.\end{aligned} \tag{3.38}$$

In (3.36) and (3.38) we have set

$$G(t)^*A(t)^* := -[[-A(t)]^\alpha G(t)]^*[-A(t)^*]^{1-\alpha}; \tag{3.39}$$

as both operators $P(t)$, $U(r, t)^*$ (with $r > t$) have their range contained in $D_{[-A(t)^*]^\eta}$ for each $\eta \in [0, 1[$, both (3.36) and (3.38) are meaningful.

Proof. Recalling (3.39), set

$$\hat{u}_0(t) := N(t)^{-1}G(t)^*A(t)^*P(t)\Phi(t, 0)Y_0; \tag{3.40}$$

note that $\hat{u}_0 \in L^2(0, T; U)$ because of (3.23) (since $2(1-\alpha-2\beta) < 1$) and Lemma 3.10 (with $J = [0, T]$). Now let $\hat{y}_0(\cdot)$ be the function (3.1) corresponding to $\hat{u}_0(\cdot)$. Then $\hat{y}_0 \in L^2(0, T; H)$ by Lemma 3.3(i). Moreover, comparing (3.1) with (3.20), and taking into account (3.16), we see that (3.37) holds. Consequently (3.40) implies (3.36). In addition, evaluating $P(0)Y_0$ by means of (3.21) with $\bar{T} = T$, we easily check that (3.35) also holds. Next, setting $p(t) := P(t)\hat{Y}_0(t)$, (3.40) and (3.37) immediately yield

$$\hat{U}_0(t) = N(t)^{-1}G(t)^*A(t)^*p(t);$$

on the other hand, by (3.37) and (3.22) with $\bar{T} = T$ we obtain the last equation in (3.38), so that the pair $(\hat{u}_0, \hat{y}_0)$ satisfies (3.38).

In order to conclude the proof of the theorem, it is sufficient to show that:

(a) If $(\hat{u}_0, \hat{y}_0)$ is a solution of the system (3.38) in $L^2(0, T; U) \times L^2(0, T; H)$, then $\hat{U}_0$ is an optimal control;

(b) The optimal control is unique.

From the equality

$$(z_1 \mid z_1) - (z_2 \mid z_2) = (z_1 - z_2 \mid z_1 - z_2) + 2\operatorname{Re}(z_2 \mid z_1 - z_2),$$

which holds true for any inner product, we derive for each $u \in L^2(0, T; U)$, denoting by $y(\cdot)$ the corresponding function (3.1)

$$J(u) - J(\hat{u}_0) = I_1(u, \hat{u}_0) + I_2(u, \hat{u}_0), \tag{3.41}$$

where $J(\cdot)$ is the cost functional appearing in (3.10) and

$$\begin{aligned} I_1(u, \hat{u}_0) := \int_0^T &\{(M(t)[y(t) - \hat{Y}_0(t)] \mid y(t) - \hat{Y}_0(t))_H \\ &+ (N(t)[u(t) - \hat{u}_0(t)] \mid u(t) - \hat{u}_0(t))_U\}\, dt \\ &+ (P_T[y(T) - \hat{Y}_0(T)] \mid y(T) - \hat{Y}_0(T))_H, \\ I_2(u, \hat{u}_0) := 2\operatorname{Re} \int_0^T &\{(M(t)\hat{Y}_0(t) \mid y(t) - \hat{Y}_0(t))_H + (N(t)\hat{u}_0(t) \mid u(t) - \hat{u}_0(t))_U\}\, dt \\ &+ 2\operatorname{Re}(P_T\hat{Y}_0(T) \mid y(T) - \hat{Y}_0(T))_H. \end{aligned}$$

Now, using (3.1) and integrating by parts,

$$\begin{aligned} I_2(u, \hat{u}_0) = 2\operatorname{Re} \int_0^T \Big\{ -\int_s^T &(M(t)\hat{Y}_0(t) \mid U(t, s)A(s)G(s)[u(s) - \hat{u}_0(s)])_H\, dt \\ &+ (N(s)\hat{u}_0(s) \mid u(s) - \hat{u}_0(s))_U \\ &- (P_T\hat{Y}_0(T) \mid U(T, s)A(s)G(s)[u(s) - \hat{u}_0(s)])_H \Big\}\, ds, \end{aligned}$$

and by the last two identities in (3.38) we easily get

$$I_2(u, \hat{u}_0) = 2 \operatorname{Re} \int_0^T (-G(s)^*A(s)^*p(s) + N(s)\hat{u}_0(s) \,|\, u(s) - \hat{u}_0(s))_U \, ds = 0.$$

On the other hand, clearly, $I_1(u, \hat{u}_0) \geqq 0$, so that (3.41) yields

$$J(u) \geqq J(\hat{u}_0) \quad \forall u \in L^2(0, T; U),$$

i.e., $\hat{u}_0$ is an optimal control. This proves (a).

Finally, if $\bar{u}$ is another optimal control, the equality $J(\hat{u}_0) = J(\bar{u})$ implies

$$I_1(\bar{u}, \hat{u}_0) = 0,$$

and by the uniform coerciveness of $N(t)$ (see (3.12)) we obtain $\bar{u} = \hat{u}_0$. This proves (b). The proof of Theorem 3.14 is complete. □

REFERENCES

[Ac] P. Acquistapace, *Evolution operators and strong solutions of abstract linear parabolic equations*, Differential and Integral Equations, 1 (1988), pp. 433-457.

[AT1] P. Acquistapace and B. Terreni, *A unified approach to abstract linear non-autonomous parabolic equations*, Rend. Sem. Mat. Univ. Padova, 78 (1987), pp. 47-107.

[AT2] ———, *On fundamental solutions for abstract parabolic equations*, in Differential Equations in Banach Spaces, Proc. Bologna 1985, A. Favini and E. Obrecht, eds., Lecture Notes in Mathematics, Vol. 1223, Springer-Verlag, Berlin, Heidelberg, 1986, pp. 1-11.

[AT3] ———, *On quasilinear parabolic systems*, Math. Ann., 282 (1988), pp. 315-335.

[ADN] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary of solutions of elliptic partial differential equations satisfying general boundary conditions*, II, Comm. Pure Appl. Math., 17 (1964), pp. 35-92.

[Am] H. Amann, *Existence and regularity for semilinear parabolic evolution equations*, Ann. Scuola Norm. Sup. Pisa, Cl. Sci. (4), 11 (1984), pp. 593-676.

[B1] V. A. Balakrishnan, *Boundary control of parabolic equations:* L-Q-R *theory*, in Theory of Nonlinear Operators, Proc. 5th Internat. Summer School, Berlin 1977, R. Kluge, ed., Abh. Akad. Wiss. DDR, Abh. Math. Naturwiss. Tech., 6N (1978), pp. 11-23.

[B2] ———, *Applied Functional Analysis*, Springer-Verlag, Berlin, Heidelberg, New York, 1981.

[D] 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, *Riccati equations with unbounded coefficients*, Ann. Mat. Pura Appl., 140 (1985), pp. 209-221.

[DI2] ———, *Bounded solutions on the real line to non-autonomous Riccati equations*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Natur. (8), 79 (1985), pp. 107-112.

[DLT] G. Da Prato, I. Lasiecka, and R. Triggiani, *A direct study of the Riccati equation arising in hyperbolic boundary control problems*, J. Differential Equations, 64 (1986), pp. 26-47.

[DS] M. C. Delfour and M. Sorine, *The linear quadratic optimal control problem for parabolic systems with boundary control through a Dirichlet condition*, in Control of Distributed Parameter Systems, Proc. 3rd IFAC Symposium, Toulouse 1982, J. P. Babary and L. Le Letty eds., Pergamon Press, Oxford, 1983, pp. 87-90.

[Fa] H. O. Fattorini, *Boundary control systems*, SIAM J. Control, 6 (1968), pp. 349-385.

[F1] F. Flandoli, *Riccati equation arising in a boundary control problem with distributed parameters*, SIAM J. Control Optim., 22 (1984), pp. 76-86.

[F2] ———, *Algebraic Riccati equation arising in boundary control problems*, SIAM J. Control Optim., 25 (1987), pp. 612-636.

[GG] G. Geymonat and P. Grisvard, *Alcuni risultati di teoria spettrale per i problemi ai limiti lineari ellittici*, Rend. Sem. Mat. Univ. Padova, 38 (1967), pp. 121-173.

[Gi] J. S. Gibson, *The Riccati integral equations for optimal control problems on Hilbert spaces*, SIAM J. Control Optim., 17 (1979), pp. 537-565.

[GT] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, Heidelberg, New York, 1977.

[Gr] P. Grisvard, *Équations différentielles abstraites*, Ann. Sci. École Norm. Sup. (4), 2 (1969), pp. 311-395.

[HLP] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, 1934.

[I] A. Ichikawa, *Filtering and control of stochastic differential equations with unbounded coefficients*, Stochastic Anal. Appl., 4 (1986), pp. 187-212.

[La] I. Lasiecka, *Unified theory for abstract parabolic boundary problems—a semigroup approach*, Appl. Math. Optim., 6 (1980), pp. 287-383.

[LT1] I. Lasiecka and R. Triggiani, *Dirichlet boundary control problem for parabolic equations with quadratic cost: analyticity and Riccati's feedback synthesis*, SIAM J. Control Optim., 21 (1983), pp. 41-67.

[LT2] ———, *Stabilization and structural assignment of Dirichlet boundary feedback parabolic equations*, SIAM J. Control Optim., 21 (1983), pp. 766-803.

[L1] J.-L. Lions, *Espaces d'interpolation et domaines des puissances fractionnaires d'opérateurs*, J. Math. Soc. Japan, 14 (1962), pp. 233-242.

[L2] ———, *Optimal Control of Systems Governed by Partial Differen'ial Equations*, Springer-Verlag, Berlin, Heidelberg, 1971.

[LM1] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes*, I, Dunod, Paris, 1968.

[LP] J.-L. Lions and J. Peetre, *Sur une classe d'espaces d'interpolation*, Inst. Hautes Études Sci. Publ. Math., 19 (1964), pp. 5-68.

[S1] M. Sorine, *Un resultat d'existence et unicité pour l'équation de Riccati stationnaire*, Rapport CRMA no. 984, University of Montreal, Montreal, Quebec, Canada, 1980.

[S2] ———, *Sur le semigroupe non linéaire associé a l'équation de Riccati*, Rapport CRMA no. 1055, University Montreal, Montreal, Quebec, Canada, 1981.

[Te] B. Terreni, *Non-homogeneous initial-boundary value problems for linear parabolic systems*, Studia Math., 92 (1989), pp. 141-175.

[Tr] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, New York, Oxford, 1978.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 1, pp. 119-138, January 1991

006

APPLICATIONS OF A SPLITTING ALGORITHM TO DECOMPOSITION IN CONVEX PROGRAMMING AND VARIATIONAL INEQUALITIES*

PAUL TSENG†

Abstract. Recently Han and Lou proposed a highly parallelizable decomposition algorithm for minimizing a strongly convex cost over the intersection of closed convex sets. It is shown that their algorithm is in fact a special case of a splitting algorithm analyzed by Gabay for finding a zero of the sum of two maximal monotone operators. Gabay's convergence analysis for the splitting algorithm is sharpened, and new applications of this algorithm to variational inequalities, convex programming, and the solution of linear complementarity problems are proposed. For convex programs with a certain separable structure, a multiplier method that is closely related to the alternating direction method of multipliers of Gabay-Mercier and of Glowinski-Marrocco, but which uses both ordinary and augmented Lagrangians, is obtained.

Key words. maximal monotone operator, augmented Lagrangian, alternating minimization

AMS(MOS) subject classifications. 49, 90

1. Introduction. One of the most important applications of convex duality theory is in decomposition algorithms for solving problems with special structure. A canonical example is the following separable convex programming problem

$$\begin{array}{ll} \text{minimize} & f(x)+g(z) \\ \text{subject to} & Ax+Bz=b, \end{array} \tag{1.1}$$

where $f:\mathcal{R}^n\to(-\infty,\infty]$ and $g:\mathcal{R}^m\to(-\infty,\infty]$ are given convex functions, A is a given $r\times n$ matrix, B is a given $r\times m$ matrix, and b is a given vector in $\mathcal{R}^r$. In our notation, all vectors are column vectors and $\mathcal{R}^k$ denotes the k-dimensional Euclidean space.

By attaching a Lagrange multiplier vector $p\in\mathcal{R}^r$ to the constraints in (1.1), the problem (1.1) can be decomposed into two independent problems involving, respectively, x and z. One algorithm based on this dual approach, proposed by Uzawa [61] and others, operates by successively minimizing the Lagrangian function

$$L(x,z,p)=f(x)+g(z)+\langle p, b-Ax-Bz\rangle,$$

with respect to x and z (with p fixed) and then updating the multipliers by the iteration

$$p:=p+c(b-Ax-Bz),$$

where c is a positive stepsize and $\langle\cdot,\cdot\rangle$ denotes the usual Euclidean inner product. (We assume for the sake of discussion that the minimum above is attained.) It can be shown that this algorithm is convergent if both f and g are strongly convex and c is chosen to be sufficiently small. (In this case the dual functional defined by $q(p)=\min_{x,z}L(x,z,p)$ is differentiable and this algorithm can be viewed as a gradient method for maximizing q.)

Unfortunately, for many problems of interest, the function f may be strongly convex but not g. This is particularly the case when a problem is transformed in a way to bring about a structure that is favorable for decomposition (see § 4 for an example).

* Received by the editors November 21, 1988; accepted for publication (in revised form) January 18, 1990.

† Laboratory for Information and Decision Systems, Room 35-205, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. This research was partially supported by United States Army Research Office contract DAAL03-86-K-0171 (Center for Intelligent Control Systems), and by National Science Foundation grant NSF-ECS-8519058.

A solution to this difficulty is suggested by a recent work of Han and Lou. In [28] they proposed a decomposition algorithm for minimizing a strongly convex function over the intersection of a finite collection of closed convex sets. It can be shown, by introducing auxiliary variables, that this convex program is a special case of (1.1) (see § 4). Moreover, it can be shown that their algorithm is similar to the dual gradient method above, except for the key difference that the Lagrangian function is replaced by an augmented Lagrangian function when the minimization is taken with respect to z.

The above discussion suggests the following generalization of the Han and Lou algorithm for solving the general problem (1.1). (The main interest here is in problems where f is strongly convex and separable but g is not strongly convex.) At each iteration of this algorithm, the ordinary Lagrangian $L(x, z, p)$ is first minimized with respect to x (with z and p held fixed), and then the augmented Lagrangian

$$L_c(x, z, p) = L(x, z, p) + c\|Ax + Bz - b\|^2/2$$

is minimized with respect to z (with x and p held fixed), where $\|\cdot\|$ denotes the norm induced by $\langle \cdot, \cdot \rangle$, i.e., $\|x\| = \sqrt{\langle x, x\rangle}$. Finally, the multipliers are updated according to the usual augmented Lagrangian iteration

$$p := p + c(b - Ax - Bz),$$

and the process is repeated. This algorithm, which for ease of reference will be called the *alternating minimization algorithm*, has the nice feature that, if B has full column rank, then both minimizations involve strongly convex objective functions. Moreover, if f is separable (in addition to being strongly convex), then the first minimization is also separable—a feature that makes this algorithm particularly suitable for problems where f is separable and g is such that the minimization of the augmented Lagrangian with respect to z is easily carried out.

The above approach of introducing a form of regularization to induce positive curvature in the objective functions is not new. It was first employed in the *alternating direction method of multipliers*, proposed by Gabay and Mercier [20], Glowinski and Marrocco [25] and extended by Gabay [17] (also see [4], [13], [14], [16], [18], [22], [23], [58] for related works), which is another multiplier method that alternates between minimization with respect to x and minimization with respect to z. The only difference between this algorithm and the one above (i.e., the alternating minimization algorithm) is that, at each iteration, x is updated by minimizing the augmented Lagrangian rather than the ordinary Lagrangian as in the above algorithm. The quadratic term of the augmented Lagrangian adversely affects the decomposition of the minimization with respect to x based on separability properties of f, and this is an advantage for the above algorithm. On the other hand, in contrast with the alternating direction method of multipliers, the penalty parameter c in the above algorithm must be chosen from a restricted range (as will be seen later), usually through trial and error.

It turns out that the alternating minimization algorithm is itself a dual application of an algorithm, suggested by Lions and Mercier and Passty and studied extensively by Gabay, for finding a zero of the sum of two maximal monotone operators. (Such operators have been studied extensively owing to their role in convex analysis and in solving certain partial differential equations. Finding a zero of the sum of these operators is a fundamental problem (see for example [5], [11], [12], [34], [50], [57]).) Let $\mathcal{H}$ be a real Hilbert space and let $\Pi : \mathcal{H} \to \mathcal{H}$ and $\Psi : \mathcal{H} \to \mathcal{H}$ be two maximal monotone operators such that Π^{-1} is in addition strongly monotone. We associate with Π and Ψ

the following problem:

$$\text{(1.2)} \qquad \text{Find } p^* \in \mathcal{H} \text{ satisfying } 0 \in \Pi(p^*) + \Psi(p^*).$$

Lions and Mercier [34] and, independently, Passty [51] suggested the following *splitting* iteration for solving (1.2), whereby a *forward Euler* step for Π is alternated with a *backward Euler* step for Ψ, i.e.,

$$\text{(1.3)} \qquad p := [I + c\Psi]^{-1}[I - c\Pi]p,$$

with c being some positive stepsize. Passty showed that, under certain assumptions, the *weighted average* of the iterates generated by (1.3), weighted by the respective stepsizes, converges to a solution of (1.2). (Interestingly, Passty's result does not require Π^{-1} to be strongly monotone.) The first "practical" convergence result was given by Gabay [19] (also see [17, Chap. 6]), who showed that the iterates converge weakly to a solution of (1.2) if c is fixed and is strictly less than twice the modulus of Π^{-1}. Gabay also gave sufficient conditions for the strong convergence of the iterates and discussed applications of the splitting iteration to decomposition in convex programming and variational inequalities. However, Gabay's applications were to the original problem (giving rise to methods such as that of Bruck [6] and of Goldstein [26]) rather than to the dual. Finally, we remark that Lemaire [32] recently gave a detailed convergence analysis for (1.3) when Ψ is the subdifferential of a closed convex function and Π is in addition strongly monotone.

The purpose of this paper is twofold: (1) to show that the alternating minimization algorithm (and hence the algorithm of Han and Lou) is an application of the iteration (1.3) to the dual of (1.1) and, (2) to give additional dual applications of (1.3) to decomposition in convex programming and variational inequalities. We begin by giving in § 2 a proof of convergence for the iteration (1.3)—different from one given by Gabay—that does not require the stepsize c to be fixed and shows that, if either Π or Ψ is strongly monotone, then the convergence is at least linear. In § 3 we apply (1.3) to variational inequalities possessing a certain separable structure to obtain a new decomposition algorithm for this problem. The latter algorithm in turn is applied, in § 4, to (1.1) to obtain the alternating minimization algorithm and, in § 5, to linear complementarity problems with positive semi-definite matrices to obtain a new matrix splitting algorithm for these problems.

We briefly describe the notation used throughout this paper. For any real matrix E, we denote by E^T the transpose of E and by $\|E\|$ the L_2-norm of E, i.e., $\|E\|$ is the square root of the largest eigenvalue of E^TE. For any set Ω, we denote by $\delta_\Omega(\cdot)$ the *indicator* function for Ω, i.e., $\delta_\Omega(x)$ is zero if $x \in \Omega$ and is ∞ otherwise. For any real Hilbert space $\mathcal{H}$ endowed with an inner product $\langle \cdot, \cdot \rangle$, we say that a multifunction $T: \mathcal{H} \to \mathcal{H}$ is a *monotone operator* if

$$\langle y - y', x - x' \rangle \geqq 0 \quad \text{whenever } y \in T(x),\ y' \in T(x').$$

It is said to be *maximal monotone* if, in addition, the graph

$$\{(x, y) \in \mathcal{H} \times \mathcal{H} \mid y \in T(x)\}$$

is not properly contained in the graph of any other monotone operator $T': \mathcal{H} \to \mathcal{H}$. We denote by $\mathcal{D}(T)$ the *effective domain* of T, i.e.,

$$\mathcal{D}(T) = \{x \in \mathcal{H} \mid T(x) \neq \emptyset\},$$

and by T^{-1} the *inverse* of T, i.e.,

$$(T^{-1})(y) = \{x \in \mathcal{H} \mid y \in T(x)\}, \qquad \forall y \in \mathcal{H}.$$

It is easily seen from symmetry that the inverse of a maximal monotone operator is also a maximal monotone operator. For any monotone operator $T: \mathcal{H} \to \mathcal{H}$, the *modulus* of T will denote the largest nonnegative scalar σ such that

$$(1.4) \qquad \langle y-y', x-x'\rangle \geq \sigma\|x-x'\|^2 \quad \text{whenever } y \in T(x), y' \in T(x'),$$

where $\|\cdot\|$ is the norm induced by $\langle \cdot, \cdot \rangle$, i.e., $\|x\| = \sqrt{\langle x, x\rangle}$. We will say that T is *strongly monotone* (or *coercive*) if its modulus is positive. For any closed convex function $h: \mathcal{H} \to (-\infty, \infty]$ and any $x \in \mathcal{H}$, we denote by dom (h) the effective domain of h, i.e.,

$$\text{dom}\,(h) = \{x \in \mathcal{H} \mid h(x) < \infty\},$$

and by $\partial h(x)$ the *subdifferential* of h at x, i.e.,

$$\partial h(x) = \{y \in \mathcal{H} \mid h(x') - h(x) \geqq \langle y, x'-x\rangle \text{ for all } x' \in \mathcal{H}\}.$$

For any other real Hilbert space $\mathcal{V}$ endowed with an inner product $\langle \cdot, \cdot \rangle$ and any continuous linear operator $A: \mathcal{V} \to \mathcal{H}$ (see [62]), we will denote by $A': \mathcal{H} \to \mathcal{V}$ the *adjoint* of A, i.e.,

$$\langle Ax, y\rangle = \langle x, A'y\rangle, \quad \forall x \in \mathcal{V}, y \in \mathcal{H},$$

and by $\|A\|$ the *operator norm* of A, i.e,

$$\|A\| = \sup_{\|x\| \leqq 1} \|Ax\|.$$

2. A splitting algorithm for the sum of two maximal monotone operators. Let $\mathcal{H}$, $\mathcal{V}$, and $\mathcal{W}$ be three real Hilbert spaces each equipped with an inner product. For convenience, we will denote each of these inner products generically as $\langle \cdot, \cdot \rangle$ and the norm induced by it as $\|\cdot\|$, with the choice of the inner product being implied by the context. Let $\Phi: \mathcal{V} \to \mathcal{V}$ and $\Gamma: \mathcal{W} \to \mathcal{W}$ be two maximal monotone operators, let $A: \mathcal{V} \to \mathcal{H}$ and $B: \mathcal{W} \to \mathcal{H}$ be two continuous linear operators, and let b be an element of $\mathcal{H}$. Consider the following problem [cf. (1.2)]:

$$(2.1) \qquad \text{Find } p^* \in \mathcal{H} \text{ satisfying } b \in A\Phi(A'p^*) + B\Gamma(B'p^*).$$

We make the following standing assumptions regarding (2.1):

Assumption A.

(a) (2.1) has a solution.

(b) Φ^{-1} is strongly monotone with modulus σ.

(c) $B\Gamma B'$ is a maximal monotone operator.

Notice that Assumption A(b) implies that Φ^{-1} is surjective (see [5, Corollary 2.4]), so that for every $p \in \mathcal{H}$ there exists an $x \in \mathcal{V}$ satisfying $x \in \Phi(A'p)$. Moreover, there exists an $x^* \in \mathcal{V}$ satisfying

$$(2.2) \qquad x^* \in \Phi(A'p^*), \quad \forall \text{ solutions } p^* \text{ of } (2.1).$$

To see the latter, note that if p_1 and p_2 are two solutions of (2.1), then there exist $x_1 \in \Phi(A'p_1)$, $z_1 \in \Gamma(B'p_1)$ such that $b = Ax_1 + Bz_1$ and there exist $x_2 \in \Phi(A'p_2)$, $z_2 \in \Gamma(B'p_2)$ such that $b = Ax_2 + Bz_2$. Hence,

$$0 = \langle x_2 - x_1, A'p_2 - A'p_1\rangle + \langle z_2 - z_1, B'p_2 - B'p_1\rangle \geqq \sigma\|x_2 - x_1\|^2,$$

where the inequality follows from the facts that Γ is monotone and that Φ^{-1} is monotone with modulus σ. Since $\sigma > 0$, this shows $x_1 = x_2$.

Consider the following algorithm for solving (2.1): Begin with any $p(0)\in\mathcal{H}$. At the tth iteration, we are given a $p(t)\in\mathcal{H}$; we generate the next iterate $p(t+1)\in\mathcal{H}$ by first computing an

$$x(t)\in\Phi(A'p(t)), \tag{2.3a}$$

then a $z(t)\in\mathcal{W}$ satisfying

$$z(t)\in\Gamma(B'[p(t)-c(t)(Ax(t)+Bz(t)-b)]), \tag{2.3b}$$

and finally setting

$$p(t+1)=p(t)+c(t)(b-Ax(t)-Bz(t)), \tag{2.3c}$$

where $c(t)$ is some positive stepsize to be specified. We will show later that $z(t)$ is well defined [i.e., there exists a $z(t)$ satisfying (2.3b)].

To see the connection between the above iteration (2.3a)-(2.3c) and the splitting iteration (1.3), let us apply B to both sides of (2.3b). By using (2.3c), we then obtain

$$[p(t)-p(t+1)]/c(t)+b-Ax(t)\in B\Gamma(B'p(t+1)),$$

which, when combined with (2.3a), yields

$$p(t+1)=[I+c(t)B\Gamma B']^{-1}(c(t)b+[I-c(t)A\Phi A']p(t)), \tag{2.4}$$

or, equivalently,

$$p(t+1)=[I+c(t)\Psi]^{-1}[I-c(t)\Pi]p(t), \tag{2.5}$$

where we let $\Pi(p)=A\Phi(A'p)-b$ and $\Psi(p)=B\Gamma(B'p)$. The iteration (2.5) is clearly of the form (1.3). In addition, Ψ is maximal monotone [cf. Assumption A(c)] and Π^{-1} is maximal strongly monotone (cf. [19, Prop. 4.1] and the maximal strongly monotone property of Φ^{-1}). A simple calculation shows that the modulus of Π^{-1} is equal to $\sigma/\|A\|^2$.

Now we show that $z(t)$ is well defined. Since $B\Gamma B'$ is by assumption a maximal monotone operator, then by a result of Minty [41], the proximal mapping $[I+c(t)B\Gamma B']^{-1}$ is single valued and defined on all of $\mathcal{H}$, so that by (2.4), $p(t+1)$ is well defined. This in turn shows that $\Gamma(B'p(t+1))$ is nonempty so that, by (2.3b) and (2.3c), there exists $z(t)\in\mathcal{W}$ satisfying (2.3b).

By using the relation (2.5), we can conclude from a result of Gabay [19, Thm. 6.1] (also see [17, Chap. 6]) that $\{p(t)\}$ converges weakly to a solution of (2.1), provided that $c(t)$ is fixed at some value between 0 and $2\sigma/\|A\|^2$ (twice the modulus of Π^{-1}) for all t. Below we sharpen this result of Gabay by showing that $\{p(t)\}$ converges weakly even if $c(t)$ is changing with t. Moreover, we show that both $\{x(t)\}$ and $\{Bz(t)\}$ converge strongly, and if either $A\Phi A'$ or $B\Gamma B'$ is strongly monotone, then the convergence is at least linear.

PROPOSITION 1. *The sequences* $\{x(t)\}$, $\{z(t)\}$, $\{p(t)\}$ *generated by* (2.3a)-(2.3c) *are well defined. If in addition* $\{c(t)\}$ *satisfies*

$$\varepsilon\leqq c(t)\leqq 2\sigma/\|A\|^2-\varepsilon,\quad \forall t, \tag{2.6}$$

for some $\varepsilon\in(0,\sigma/\|A\|^2]$, *then the following hold*:

(a) $\{x(t)\}\to x^*$ *in the strong topology.*

(b) $\{Bz(t)\}\to b-Ax^*$ *in the strong topology.*

(c) $\{p(t)\}\to$ *a solution of* (2.1) *in the weak topology.*

(d) *If either* $A\Phi A'$ *or* $B\Gamma B'$ *is strongly monotone, then* $\{x(t)\}$, $\{Bz(t)\}$, $\{p(t)\}$ *converge at least linearly with a convergence ratio of* $\sqrt{(1-\delta^2\varepsilon^2)/(1+\eta^2\varepsilon^2)}$, *where* δ *and* η *denote the modulus of, respectively,* $A\Phi A'$ *and* $B\Gamma B'$.

The proof of Proposition 1, which is based on an argument used by Glowinski and Le Tallec [23] (also see [4, § 3.4.4]) for the alternating direction method of multipliers, is given in Appendix A. It is unclear if the proof given by Gabay can be extended to show Proposition 1 since it uses a lemma of Opial [46] which requires the algorithmic mapping to be *stationary.*

Note that Proposition 1(b) implies that if $B'B$ is an isomorphism of $\mathscr{H}$, then $\{z(t)\}$ converges. Also, it can be shown that Proposition 1 holds even if the solutions of (2.3a) and (2.3b) are computed inexactly. Unfortunately, the amount of inexactness allowable cannot be easily estimated. Also, notice that, by letting $\Pi(p)=A\Phi(A'p)-b$ and $\Psi(p)=B\Gamma(B'p)$, we can restrict our attention to iterations of the simpler form (2.5). However, by doing so, we will only be able to infer the convergence of $\{Ax(t)\}$, not the convergence of $\{x(t)\}$.

We have thus far assumed $B\Gamma B'$ to be maximal (it is automatically monotone since Γ is monotone), but for practical uses we must translate this assumption into conditions on Γ and B that are more easily verified. For this purpose, the following equivalent set of conditions given by Gabay [19, proof of Prop. 4.1] (also see [17, Chap. 6]) will be useful to us.

LEMMA 1. *$B\Gamma B'$ is maximal monotone if and only if, for each $c>0$, the effective domain of the multifunction $p\to(\Gamma^{-1}+cB'B)^{-1}B'p$ is all of $\mathscr{H}$.*

By Lemma 1, $B\Gamma B'$ is maximal monotone if, for each $c>0$, $\Gamma^{-1}+cB'B$ is strongly monotone (since $\Gamma^{-1}+cB'B$ is then surjective by [5, Corollary 2.4]), and the latter holds if either $B'B$ is an isomorphism of $\mathscr{H}$ or if Γ^{-1} is strongly monotone (cf. (4.7) and (4.8) in [19]). These requirements can, in certain special cases, be relaxed further. For example, if Γ^{-1} is the subdifferential mapping of a closed proper convex function g, then it suffices to require that the function $z\to g(z)+\|Bz\|^2$ attains its minimum at some point (see § 4).

3. Application to variational inequalities. Let $\mathscr{X}$ and $\mathscr{Z}$ be two polyhedral sets in, respectively, $\mathscr{R}^n$ and $\mathscr{R}^m$. Let $R:\mathscr{R}^n\to\mathscr{R}^n$ and $S:\mathscr{R}^m\to\mathscr{R}^m$ be two single valued continuous functions and let $f:\mathscr{R}^n\to(-\infty,\infty]$ and $g:\mathscr{R}^m\to(-\infty,\infty]$ be two closed convex functions. Also, let A be an $r\times n$ matrix, B be an $r\times m$ matrix, and b be an element of $\mathscr{R}^r$. Consider the following problem:

$$\text{Find } x^*\in\mathscr{X} \text{ and } z^*\in\mathscr{Z} \text{ satisfying } Ax^*+Bz^*=b \text{ and}$$

$$\langle x-x^*, R(x^*)\rangle+\langle z-z^*, S(z^*)\rangle+f(x)-f(x^*)+g(z)-g(z^*)\geqq 0, \tag{3.1}$$

$$\forall x\in\mathscr{X},\ z\in\mathscr{Z} \text{ satisfying } Ax+Bz=b.$$

This problem, called the *variational inequality* problem, has numerous applications to numerical computation—including the solution of a system of equations, constrained and unconstrained optimization, traffic assignment problems, game theory, and saddle point problems (see [1], [4, § 3.5], [9], [16], [22], [30]). For example, the convex program (1.1) is a special case of (3.1) with R and S taken to be the zero functions. In this section we will derive a decomposition algorithm for (3.1) by applying the splitting iteration (2.3a)-(2.3c).

We make the following assumptions regarding (3.1).

Assumption B.

(a) (3.1) has a solution.

(b) $R+\partial f$ is strongly monotone (with modulus σ).

(c) For each $c>0$, the effective domain of the multifunction $p\to(S+\partial g+\partial\delta_{\mathscr{Z}}+cB^TB)^{-1}B^Tp$ is all of $\mathscr{R}^r$.

(d) There exist $x \in \mathrm{ri}\,(\mathrm{dom}\,(f)) \cap \mathscr{X}$ and $z \in \mathrm{ri}\,(\mathrm{dom}\,(g)) \cap \mathscr{Z}$ satisfying $Ax + Bz = b$.

Part (c) of Assumption B seems a bit complicated, but notice that it holds automatically if either $S + \partial g$ is strongly monotone or if B has full column rank (cf. Lemma 1). Part (d) of Assumption B combines the usual feasibility assumption for (3.1) with a constraint qualification. The latter is required in order to assert the existence of an optimal Lagrange multiplier vector associated with the constraints $Ax + Bz = b$. Note that we could have embedded the constraints $x \in \mathscr{X}$ and $z \in \mathscr{Z}$ in, respectively, f and g by the use of the indicator functions $\delta_{\mathscr{X}}$ and $\delta_{\mathscr{Z}}$. But had we done so, we would not have been able to infer (from the available duality results) the existence of an optimal Lagrange multiplier vector without making additional assumptions.

Since $R + \partial f$ is strongly monotone, it is easily seen that the first n coordinates of a solution of (3.1) are uniquely determined; i.e., there exists an $x^* \in \mathscr{X}$ such that every solution of (3.1) is of the form (x^*, z^*), where z^* is some element of $\mathscr{Z}$.

We claim that (3.1) is a special case of the problem (2.1). To see this, let (x^*, z^*) be a solution of the variational inequality (3.1). Then, it is easily seen that (x^*, z^*) solves the following convex program:

$$\text{(3.2)} \qquad \begin{array}{ll} \text{minimize} & \langle R(x^*), x\rangle + \langle S(z^*), z\rangle + f(x) + g(z) \\ \text{subject to} & x \in \mathscr{X},\ z \in \mathscr{Z},\ Ax + Bz = b, \end{array}$$

so that, by [54, Thm. 28.2] (also using Assumption B(d)), there exists an optimal Lagrange multiplier vector $p^* \in \mathscr{R}^r$ associated with the constraints $Ax + Bz = b$. From the Kuhn-Tucker conditions for (3.2) we then have

$$\text{(3.3a)} \qquad A^T p^* \in R(x^*) + \partial f(x^*) + \partial \delta_{\mathscr{X}}(x^*),$$

$$\text{(3.3b)} \qquad B^T p^* \in S(z^*) + \partial g(z^*) + \partial \delta_{\mathscr{Z}}(z^*),$$

$$\text{(3.3c)} \qquad Ax^* + Bz^* = b.$$

Let $F: \mathscr{R}^n \to \mathscr{R}^n$ and $G: \mathscr{R}^m \to \mathscr{R}^m$ be the multifunctions given by, respectively,

$$\text{(3.4a)} \qquad F(x) = R(x) + \partial f(x) + \partial \delta_{\mathscr{X}}(x),$$

$$\text{(3.4b)} \qquad G(z) = S(z) + \partial g(z) + \partial \delta_{\mathscr{Z}}(z).$$

Then, we can write (3.3a)-(3.3c) equivalently as

$$\text{(3.5)} \qquad b \in AF^{-1}(A^T p^*) + BG^{-1}(B^T p^*).$$

Since both $\partial f + \partial \delta_{\mathscr{X}}$ and $\partial g + \partial \delta_{\mathscr{Z}}$ are maximal monotone (see Rockafellar [55]) and both R and S are monotone and continuous, it is easily seen that both F and G are maximal monotone. (In general, the sum of two maximal monotone operators need not be maximal monotone.) Since F is, in addition, strongly monotone with modulus σ [cf. Assumption B(b)], it then follows that the problem of finding a p^* satisfying (3.5) is a special case of (2.1), and moreover, parts (a) and (b) of Assumption A hold. By Assumption B(d), for each $c > 0$, the effective domain of the multifunction $p \to (G + cB^T B)^{-1} B^T p$ is all of $\mathscr{R}^r$, so that, by Lemma 1, $BG^{-1}B^T$ is maximal monotone. This shows that part (c) of Assumption A also holds.

Let us apply the splitting iteration (2.3a)-(2.3c) to the problem of finding a p^* satisfying (3.5). This generates a sequence of iterates $\{p(t)\}$, $\{x(t)\}$ and $\{z(t)\}$ according to the following iteration: Given $p(t)$, first compute an $x(t) \in \mathscr{X}$ satisfying

$$\text{(3.6a)} \qquad \langle x - x(t), R(x(t)) - A^T p(t)\rangle + f(x) - f(x(t)) \geqq 0, \quad \forall x \in \mathscr{X},$$

then compute a $z(t) \in \mathscr{Z}$ satisfying

$$\langle z - z(t), S(z(t)) - B^T(p(t) - c(t)(Ax(t) + Bz(t) - b))\rangle + g(z) - g(z(t)) \geqq 0, \quad \forall z \in \mathscr{Z}, \tag{3.6b}$$

and finally update the multiplier vector by

$$p(t+1) = p(t) + c(t)(b - Ax(t) - Bz(t)), \tag{3.6c}$$

where $c(t)$ is some positive stepsize. We leave the issue of computing $x(t)$ and $z(t)$ open. (Methods for computing $x(t)$ and $z(t)$ are discussed in [4], [10], [16], [22], [30], [49].)

Since Assumption A holds for (3.5), we immediately obtain from Proposition 1 the following proposition.

PROPOSITION 2. *The sequence of iterates* $\{x(t)\}$, $\{z(t)\}$, $\{p(t)\}$ *generated by* (3.6a)–(3.6c) *are well defined. If in addition* $\{c(t)\}$ *satisfies*

$$\varepsilon \leqq c(t) \leqq 2\sigma/\|A\|^2 - \varepsilon, \quad \forall t,$$

for some $\varepsilon \in (0, \sigma/\|A\|^2]$, *then the following hold*:

(a) $\{x(t)\} \to x^*$.

(b) $\{Bz(t)\} \to b - Ax^*$.

(c) $\{p(t)\} \to$ *a solution of* (3.5).

(d) *If either* $AF^{-1}A^T$ *or* $BG^{-1}B^T$ *is strongly monotone, where F and G are given by* (3.4a) *and* (3.4b), *respectively, then* $\{x(t)\}$, $\{Bz(t)\}$, $\{p(t)\}$ *converge at least linearly with a convergence ratio of* $\sqrt{(1-\delta^2\varepsilon^2)/(1+\eta^2\varepsilon^2)}$, *where* δ *and* η *denote the modulus of, respectively,* $AF^{-1}A^T$ *and* $BG^{-1}B^T$.

It was pointed out to us by the referee that the relationship between (3.1) and (3.5) also follows from a general duality framework studied by Gabay [19, § 3]. In particular, let $\mathscr{H}$ and $\mathscr{V}$ be two real Hilbert spaces, let Λ be a maximal monotone operator on $\mathscr{H}$, let $h: \mathscr{V} \to (-\infty, \infty]$ be a closed convex function, and let E be a continuous linear operator from $\mathscr{H}$ to $\mathscr{V}$. For each variational inequality of the form:

$$\text{Find } u \text{ such that } 0 \in \Lambda(u) + E'\partial h(Eu), \tag{3.7}$$

let us associate a dual variational inequality of the following form:

$$\text{Find } \lambda \text{ such that } 0 \in \Lambda_E^T(\lambda) + (\partial h)^{-1}(\lambda), \tag{3.8}$$

where $\Lambda_E^T(\lambda) = \{-Ev \mid -E'\lambda \in \Lambda(v)\} = -E\Lambda^{-1}(-E'\lambda)$. Gabay showed that, under the qualification condition

$$\text{there exists } u \in \text{int}\,(\mathscr{D}(\Lambda)) \text{ such that } Eu \in \text{dom}\,(h), \tag{3.9}$$

a vector u is a solution of (3.7) if and only if there exists a solution λ of (3.8) such that $-E'\lambda \in \Lambda(u)$ [19, Thm. 3.1]. Now, it is easily seen that (3.1) is a special case of (3.7) where h is the indicator function of $\{-b\}$, $\Lambda(x, z) = F(x) + G(z)$ with F and G given by (3.4a), (3.4b), respectively, and $E = [-A \;\; -B]$. Hence, we obtain from (3.8) that the corresponding dual variational inequality (in the sense of Gabay) is

$$\text{Find } \lambda \text{ such that } 0 \in AF^{-1}(A^T\lambda) + BG^{-1}(B^T\lambda) - b,$$

which is exactly (3.5). An important advantage of Gabay's duality result is that it applies to general maximal monotone operators. However, the qualification condition (3.9), when applied to (3.1), yields the condition: there exist $x \in \text{int}\,(\text{dom}\,(f) \cap \mathscr{X})$ and $z \in \text{int}\,(\text{dom}\,(g) \cap \mathscr{Z})$ satisfying $Ax + Bz = b$, which is stronger than our qualification condition, namely Assumption B(d).

4. Application to separable convex programming: the alternating minimization algorithm. Let us return to the separable convex program (1.1):

$$\text{(4.1)}\qquad \begin{aligned}&\text{minimize} && f(x)+g(z)\\ &\text{subject to} && Ax+Bz=b,\ x\in\mathscr{X},\ z\in\mathscr{Z},\end{aligned}$$

where $f:\mathscr{R}^n\to(-\infty,\infty]$, $g:\mathscr{R}^m\to(-\infty,\infty]$ are closed convex functions, A is an $r\times n$ matrix, B is an $r\times m$ matrix, and b is a vector in $\mathscr{R}^r$. In this section we will derive the alternating minimization algorithm for solving (4.1) by applying the iteration (3.6a)-(3.6c).

We make the following assumptions regarding (4.1).

Assumption C.

(a) f is strongly convex with modulus $\alpha>0$, i.e., for any $\lambda\in(0,1)$, there holds

$$\text{(4.2)}\quad \lambda f(x)+(1-\lambda)f(y)-f(\lambda x+(1-\lambda)y)\geqq\alpha\lambda(1-\lambda)\|x-y\|^2,\quad \forall x\in\mathscr{R}^n,\quad \forall y\in\mathscr{R}^n.$$

(b) The function $g(z)+\|Bz\|^2$ attains its minimum at some point in $\mathscr{Z}$.

(c) There exist $x\in\operatorname{ri}(\operatorname{dom}(f))\cap\mathscr{X}$ and $z\in\operatorname{ri}(\operatorname{dom}(g))\cap\mathscr{Z}$ satisfying $Ax+Bz=b$.

Assumption C(c) implies that the problem (4.1) is feasible. We claim that (4.1) in fact has an optimal solution. To see this, note that because f and g are closed and f is strongly convex, if (4.1) does not have an optimal solution, then there must exist z and w such that $Bw=0$ and $g(z+\lambda w)$ is strictly decreasing with $\lambda\geqq 0$—contradicting Assumption C(b). Moreover, the strict convexity of f implies that the x component of an optimal solution of (4.1) is *unique*, which we denote by x^*.

Notice that Assumption C(b) holds if either g has a minimizer or B has full column rank. If Assumption C(b) does not hold, but (4.1) has an optimal solution, then we can define the perturbation function $h(w)=\inf\{g(z)\,|\,w=Bz,\ z\in\mathscr{Z}\}$, which is proper convex. If in addition h is closed, we can instead solve the reduced problem $\min_{x,w}\{f(x)+h(w)\,|\,Ax+w=b\}$, which can be seen to satisfy parts (a) and (b) of Assumption C.

(4.1) is clearly a special case of the variational inequality problem (3.1). Furthermore, the strong convexity condition (4.2) implies that ∂f is strongly monotone with modulus 2α. Hence parts (a), (b), and (d) of Assumption B hold (with $\sigma=2\alpha$). Finally, by Assumption C(b), for any $p\in\mathscr{R}^r$ and $c>0$, the function $z\to g(z)-\langle p, Bz\rangle+c\|Bz\|^2/2$ attains its minimum at some point z. Then, z satisfies $B^Tp\in\partial g(z)+cB^TBz$, so part (b) of Assumption B holds.

By applying the iteration (3.6a)-(3.6c) to (4.1), we obtain the following algorithm, earlier named the alternating minimization algorithm, that generates a sequence of iterates $\{x(t)\}$, $\{z(t)\}$, $\{p(t)\}$ according to:

$$\text{(4.3a)}\qquad x(t)=\operatorname*{argmin}_{x\in\mathscr{X}}\{f(x)-\langle p(t), Ax\rangle\},$$

$$\text{(4.3b)}\qquad z(t)=\operatorname*{argmin}_{z\in\mathscr{Z}}\{g(z)-\langle p(t), Bz\rangle+c(t)\|Ax(t)+Bz-b\|^2/2\},$$

$$\text{(4.3c)}\qquad p(t+1)=p(t)+c(t)(b-Ax(t)-Bz(t)),$$

where each $c(t)$ is some positive stepsize. ($p(0)$ is arbitrarily chosen.)

Since Assumption B holds for this special case of (3.1), convergence of the above algorithm follows from Proposition 2.

PROPOSITION 3. *The sequences* $\{x(t)\}$, $\{z(t)\}$, $\{p(t)\}$ *generated by* (4.3a)-(4.3c) *are well defined. If in addition* $\{c(t)\}$ *satisfies*

$$\varepsilon \leqq c(t) \leqq 4\alpha/\|A\|^2 - \varepsilon, \quad \forall t,$$

for some $\varepsilon \in (0, 2\alpha/\|A\|^2]$, *then the following hold*:

(a) $\{x(t)\} \to x^*$.

(b) $\{Bz(t)\} \to b - Ax^*$.

(c) $\{p(t)\} \to$ *an optimal Lagrange multiplier vector associated with the constraints* $Ax + Bz = b$.

(d) *If either* $A(\partial f)^{-1}A^T$ *or* $B(\partial g)^{-1}B^T$ *is strongly monotone, then the rate of convergence of* $\{x(t)\}$, $\{Bz(t)\}$, $\{p(t)\}$ *is at least linear.*

(e) *If the convex function* $g(z) + \|Bz\|^2$ *has bounded level sets, then* $\{z(t)\}$ *is bounded and, for any of its limit points* z^∞, (x^*, z^∞) *is an optimal solution of* (4.1).

Proof. Parts (a)-(d) follow directly from Proposition 2. To prove part (e), let z^* be an element of $\mathscr{Z}$ which, together with x^*, forms an optimal solution of (4.1). Since $z(t) = \operatorname{argmin}_{z \in \mathscr{Z}} \{g(z) - \langle p(t+1), Bz\rangle\}$ for all t [cf. (4.3b)-(4.3c)], we then have that

$$g(z(t)) - \langle p(t+1), Bz(t)\rangle \leqq g(z^*) - \langle p(t+1), Bz^*\rangle, \quad \forall t.$$

Since $\{Bz(t)\} \to b - Ax^* = Bz^*$ and $\{p(t)\}$ is bounded [cf. parts (b) and (c)], this yields

$$\limsup_{t \to \infty} \{g(z(t))\} \leqq g(z^*). \tag{4.4}$$

Hence $g(z(t)) + \|Bz(t)\|^2$ is bounded and, by hypothesis, $\{z(t)\}$ is bounded. Since g is closed, (4.4) implies that each limit point of $\{z(t)\}$, say z^∞, satisfies $g(z^\infty) \leqq g(z^*)$. Since $Bz^\infty = b - Ax^*$ (cf. part (b)), then (x^*, z^∞) is feasible for (4.1) and its cost $f(x^*) + g(z^\infty)$ does not exceed $f(x^*) + g(z^*)$. This shows that (x^*, z^∞) is an optimal solution of (4.1). □

In practice, the threshold $4\alpha/\|A\|^2$ is typically unknown, and some trial and error may be required to select the sequence $c(t)$. This is a drawback of the alternating minimization algorithm.

Remark 1. Note that the hypothesis in Proposition 3(e) holds if B has full column rank or if g has bounded level sets. In practice, the latter can always be enforced by constraining z to be inside the ball $\{z \in \mathscr{R}^m \mid \|z\| \leqq \mu\}$ with μ taken to be a very large number. An example for which Proposition 3(d) applies is when $f(x) = \|x - d\|^2/2$ for some $d \in \mathscr{R}^n$ and A has full row rank. Straightforward calculation finds that $A(\partial f)^{-1}(A^T p) = AA^T p + Ad$ and hence $A(\partial f)^{-1}A^T$ is strongly monotone with a modulus equal to the smallest eigenvalue of AA^T.

Remark 2. In the special case where $A = -I$ and f is the indicator function for $\{0\}$, it can be seen that the alternating minimization algorithm reduces to the *method of multipliers* [27], [29], [53] (also see [3], [4], [36], [52], [56]) for minimizing $g(z)$ subject to $Bz = b$, i.e.,

$$z(t) = \operatorname{argmin}_z \{g(z) - \langle p(t), Bz\rangle + c(t)\|b - Bz\|^2/2\}, \quad p(t+1) = p(t) + c(t)(b - Bz(t)).$$

By Proposition 3 (and using the observation that the modulus of ∂f is "infinite"), if $\{c(t)\}$ is bounded away from zero, then both $\{Bz(t)\}$ and $\{p(t)\}$ converge.

Remark 3. In the special case where $B = -I$ and g is the indicator function for $\{0\}$, the alternating minimization algorithm can be seen to reduce to the dual gradient method discussed in § 1 for minimizing $f(x)$ subject to $Ax = b$, i.e.,

$$x(t) = \operatorname*{argmin}_x \{f(x) - \langle p(t), Ax\rangle\}, \qquad p(t+1) = p(t) + c(t)(b - Ax(t)).$$

By Proposition 3, if $\{c(t)\}$ is bounded strictly between zero and $4\alpha/\|A\|^2$, then both $\{x(t)\}$ and $\{p(t)\}$ converge. This method was first proposed by Uzawa [61] for the more general case where f is strictly convex, but no explicit bound on the stepsizes was given. Other discussions of this algorithm can be found in [3, § 2.6] and in [15], [31], [36], [52].

4.1 Minimization of a strongly convex function over the intersection of a finite collection of closed convex sets. Consider the following problem.

$$\text{(4.5)} \qquad \begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && x \in \mathscr{X}_1 \cap \cdots \cap \mathscr{X}_k, \end{aligned}$$

where $f: \mathscr{R}^n \to \mathscr{R}$ is a strongly convex differentiable function (with modulus, say, α) and each $\mathscr{X}_i$ is a closed convex set in $\mathscr{R}^n$. This is a well studied problem in optimization. In particular, there has been proposed a number of algorithms for solving this problem based on the splitting idea (see Lions and Temam [35] and Temam [59]). In this section we show that the recent algorithm of Han and Lou [28] for solving this problem is a special case of the alternating minimization algorithm; hence, it also uses the splitting idea. Moreover, we will use Proposition 3 to improve the convergence results given by Han and Lou.

We make the following assumption regarding (4.5).

Assumption D. Either (a) ri $(\mathscr{X}_1) \cap \cdots \cap$ ri $(\mathscr{X}_k) \neq \varnothing$ or (b) $\mathscr{X}_1 \cap \cdots \cap \mathscr{X}_k \neq \varnothing$ and all $\mathscr{X}_i$'s are polyhedral sets.

Since $\mathscr{X}_1 \cap \cdots \cap \mathscr{X}_k \neq \varnothing$ by Assumption D and f has bounded level sets (since f is strongly convex), it follows that (4.5) has an optimal solution which, by the strict convexity of f, is unique. We will denote this optimal solution by x^*.

By introducing the auxiliary variables $z_1, z_2, \ldots, z_k$, we can transform (4.5) into the following form:

$$\text{(4.6)} \qquad \begin{aligned} &\text{minimize} && f(x) + g(z_1, \ldots, z_k) \\ &\text{subject to} && x = z_i, \qquad i = 1, \cdots, k, \end{aligned}$$

where $g: \mathscr{R}^{nk} \to (-\infty, \infty]$ is the indicator function for $\mathscr{X}_1 \times \cdots \times \mathscr{X}_k$, i.e.,

$$g(z_1, \cdots, z_k) = \sum_i \delta_{\mathscr{X}_i}(z_i).$$

The problem (4.6) is clearly a special case of (4.5) where f and g are as above, b is the zero vector in $\mathscr{R}^{nk}$, B is the negative of the $kn \times kn$ identity matrix, and A is the $kn \times n$ matrix composed of $kn \times n$ identity matrices stacked one on top of the next. Moreover, since f is strongly convex, part (a) of Assumption C holds. Since the function $(z_1, \cdots, z_k) \to g(z_1, \cdots, z_k) + \sum_i \|z_i\|^2$ has bounded level sets so that it attains its minimum at some point, part (b) of Assumption C also holds. Finally, since dom $(f) = \mathscr{R}^n$ and dom $(g) = \mathscr{X}_1 \times \cdots \times \mathscr{X}_k$, Assumption D implies that part (c) of Assumption C holds.

Motivated by the above observation, let us apply the alternating minimization algorithm to the problem (4.6). This produces the following iteration:

$$\text{(4.7a)} \qquad x(t) = \operatorname*{argmin}_x \left\{ f(x) - \sum_i \langle p_i(t), x \rangle \right\},$$

$$\text{(4.7b)} \qquad z_i(t) = \operatorname*{argmin}_{z_i \in \mathscr{X}_i} \{ \langle p_i(t), z_i \rangle + c(t) \|x(t) - z_i\|^2/2 \}, \qquad i = 1, \cdots, k,$$

$$\text{(4.7c)} \qquad p_i(t+1) = p_i(t) + c(t)(z_i(t) - x(t)), \qquad i = 1, \cdots, k,$$

where p_i ($i=1,\cdots,k$) is a Lagrange multiplier vector associated with the constraints $x=z_i$ and $c(t)$ is some positive stepsize. (The initial multipliers $p_i(0)$ are chosen arbitrarily.) Note that the iterations (4.7b)-(4.7c) are highly parallelizable, and the same is true for iteration (4.7a) if f is separable.

Since Assumption C holds for this special case of (4.1), convergence of the iterates generated by (4.7a)-(4.7c) follows from Proposition 3.

COROLLARY 1. *The sequences* $\{x(t)\}$, $\{z(t)\}$, $\{p(t)\}$ *generated by* (4.7a)-(4.7c) *are well defined. If in addition* $\{c(t)\}$ *satisfies*

$$\varepsilon \leqq c(t) \leqq 4\alpha/k-\varepsilon, \quad \forall t,$$

for some $\varepsilon \in (0, 2\alpha/k]$, *then the following hold*:

(a) $\{x(t)\} \to x^*$.

(b) $\{z_i(t)\} \to x^*$, for all i.

(c) $\{p_i(t)\} \to$ *an optimal Lagrange multiplier vector associated with the constraints* $x=z_i$ *in* (4.6), *for all i.*

To see the connection between the above algorithm and that of Han and Lou, let f^* denote the conjugate function of f [54], i.e., $f^*(y)=\sup_x \{\langle y, x\rangle - f(x)\}$ for all $y \in \mathcal{R}^n$. Since f is strongly convex, then f^* is differentiable everywhere, so that (4.7a) is equivalent to

$$x(t)=\nabla f^*\left(\sum_i p_i(t)\right). \tag{4.8a}$$

Also (4.7b) can be written equivalently as

$$z_i(t)=P_{\mathcal{X}_i}(x(t)-p_i(t)/c(t)), \qquad i=1,\cdots,k, \tag{4.8b}$$

where $P_{\mathcal{X}_i}(\cdot)$ denotes the orthogonal projection onto $\mathcal{X}_i$ in the norm $\|\cdot\|$. The iteration (4.8a)-(4.8b), (4.7c) can be seen to be *identical* to that in the Han and Lou algorithm. On the other hand, our results improve upon those obtained by Han and Lou since their algorithm further restricts $p_i(0)$ to be zero for all i and $c(t)$ to take on a fixed value inside $(0, 2\alpha/k]$ for all t. In addition, Corollary 3 asserts convergence of the iterates $\{x(t)\}$, $\{z_1(t)\}, \cdots, \{z_k(t)\}$, $\{p(t)\}$ without requiring $\mathcal{X}_1 \cap \cdots \cap \mathcal{X}_k$ to have a nonempty interior (compare with [28, Thm. 4.9]). We remark that a different extension of the Han and Lou algorithm which allows inexact minimization in (4.7a), (4.7b) was recently proposed by Mouallif, Nguyen, and Strodiot [44].

5. Application to linear complementarity problems. Let M be an $r \times r$ positive semidefinite (not necessarily symmetric) matrix, i.e., $\langle p, Mp\rangle \geqq 0$ for all p. Let w be a vector in $\mathcal{R}^r$. Consider the following problem associated with M and w:

$$\text{Find } p^* \in \mathcal{R}^r \text{ satisfying } Mp^*+w \geqq 0, p^* \geqq 0, \langle Mp^*+w, p^*\rangle = 0. \tag{5.1}$$

We assume that (5.1) has a solution. This problem, called the *linear complementarity problem* (LCP for short), is a fundamental problem in optimization (see [2], [9], [37], [45]).

One method for solving (5.1) is based on the notion of *matrix splitting*. In this method, we fix a *splitting* (J, K) of M [47], i.e.,

$$M=J+K. \tag{5.2}$$

At the tth iteration, we are given an iterate $p(t) \geqq 0$ ($p(0)$ is chosen arbitrarily); we choose a *relaxation* parameter $\omega(t)>0$ and then compute the next iterate $p(t+1)$

satisfying the following linear complementarity condition

$$(5.3\text{a})\qquad (\omega(t)I+K)p(t+1)-(\omega(t)I-J)p(t)+w\geqq 0, \qquad p(t+1)\geqq 0,$$

$$(5.3\text{b})\qquad \langle(\omega(t)I+K)p(t+1)-(\omega(t)I-J)p(t)+w, p(t+1)\rangle = 0.$$

(Note that, in the absence of the nonnegativity constraints, the iteration (3.5a)-(3.5b) reduces to the iteration $(\omega(t)I+K)p(t+1)=(\omega(t)I-J)p(t)-w$, which is well known in numerical analysis (see [47]).)

Convergence of the iterates $p(t)$ generated by the matrix splitting iteration (5.3a)-(5.3b) has been well studied (see [33], [37], [48]). However, convergence typically requires that the solution is unique (in addition to other assumptions on the problem and on the matrix splitting), which does not necessarily hold for the problem (5.1). Below we apply the iteration (3.6a)-(3.6c) to the problem (5.1) to obtain a (new) splitting (J, K) of M for which the iterates generated by (5.3a)-(5.3b) *converge to a solution of* (5.1). To the best of our knowledge, this is the first matrix splitting algorithm that is provably convergent on problems having possibly multiple solutions.

PROPOSITION 4. *Let J and K be matrices satisfying* (5.2) *and suppose that*

$$(5.4\text{a})\qquad J=APA^T,$$

for some $n\times n$ $(n\leqq r)$ *positive definite* matrix P and some $r\times n$ matrix A, and

$$(5.4\text{b})\qquad K=CQC^T,$$

for some $m\times m$ $(m\leqq r)$ *positive definite* matrix Q and some $r\times m$ matrix C. If $1/\omega(t)$ is bounded strictly between zero and $2\sigma/\|A\|^2$ for all t, where σ is the smallest eigenvalue of $((P^{-1})^T+P^{-1})/2$, then $\{p(t)\}$ generated by (5.3a)-(5.3b) is well defined and converges to a solution of (5.1). If, in addition, J is positive definite, then the rate of convergence is at least linear.

Proof. The basic idea of the proof is to use (5.4a)-(5.4b) to convert (5.1) into a special case of (3.1). Then, we show that (3.6a)-(3.6c) applied to this special case is identical to (5.3a)-(5.3b).

First, we introduce a slack variable $s=Mp+w$ so that (5.1) can be written equivalently as

$$\text{Find } p^* \text{ and } s^* \text{ such that } Mp^*-s^*=-w,\ s^*\geqq 0,\ p^*\geqq 0,\ \langle s^*, p^*\rangle = 0.$$

By using the observation that the set of conditions $s\geqq 0$, $p\geqq 0$, $\langle s, p\rangle=0$ is equivalent to $-p\in\partial\delta(s)$, where δ is the indicator function for the nonnegative orthant in $\mathcal{R}^r$, we obtain, upon using (5.4a)-(5.4b), that the above problem is equivalent to

$$\text{Find } p^* \text{ and } s^* \text{ such that } APA^Tp^*+CQC^Tp^*-s^*=-w, \qquad 0\in\partial\delta(s^*)+p^*.$$

By introducing the auxiliary variables $x=PA^Tp$ and $u=QC^Tp$, we can write the above problem as

$$(5.5)\qquad \begin{aligned}&\text{Find } p^*,\ s^*,\ x^* \text{ and } u^* \text{ such that } Ax^*+Cu^*-s^*=-w,\\ &-p^*\in\partial\delta(s^*),\quad A^Tp^*=P^{-1}x^*,\quad C^Tp^*=Q^{-1}u^*,\end{aligned}$$

which in turn is equivalent to the following variational inequality problem:

$$(5.6)\qquad \begin{aligned}&\text{Find } (x^*, u^*, s^*) \text{ such that } Ax^*+Cu^*-s^*=-w,\ s^*\geqq 0, \text{ and}\\ &\langle x-x^*, P^{-1}x^*\rangle+\langle u-u^*, Q^{-1}z^*\rangle\geqq 0,\\ &\text{for all } (x, u, s) \text{ satisfying } Ax+Cu-s=-w,\ s\geqq 0.\end{aligned}$$

Hence, (5.6) has a solution, and any solution of (5.1) is an optimal Lagrange multiplier vector associated with the constraints $Ax+Cu-s=-w$.

By comparing (5.6) with (3.1), we see that (5.6) is a special case of (3.1) with A as above and with $B=[C\ -I]$, $b=-w$, $R(x)=P^{-1}x$, $S(u,s)=Q^{-1}u$, $f(x)=0$, $g(u,s)=0$, $\mathscr{X}=\mathscr{R}^n$, and $\mathscr{Z}=\mathscr{R}^m\times[0,\infty)^r$. Moreover, since (5.6) has a solution and its constraint set is polyhedral, parts (a) and (d) of Assumption B hold. Since P is positive definite, then the operator $x\to P^{-1}x$ is maximal strongly monotone (with modulus equal to the smallest eigenvalue of $((P^{-1})^T+P^{-1})/2$) and part (b) of Assumption B holds. Finally, for each $c>0$, we have from the positive definite property of Q that $\langle u, Q^{-1}u\rangle + c\|Cu-s\|^2>0$ for all $(u,s)\neq 0$. This implies that, for each $c>0$, the operator $(u,s)\to(Q^{-1}u, \partial\delta(s))+c(C^TCu-C^Ts, -Cu+s)$ is maximal strongly monotone, so that it is surjective (cf. [5, Corollary 2.4]) and part (c) of Assumption B holds.

Let us apply the iteration (3.6a)-(3.6c), with $c(t)=1/\omega(t)$, to this special case of (3.1). This generates a sequence of iterates $\{x(t)\}$, $\{u(t)\}$, $\{s(t)\}$, $\{p(t)\}$ satisfying $p(t)\geqq 0$, $s(t)\geqq 0$, and

$$\langle x-x(t), P^{-1}x(t)-A^Tp(t)\rangle\geqq 0, \quad \forall x\in\mathscr{R}^n,$$

$$\langle u-u(t), Q^{-1}u(t)-C^T(p(t)-(Ax(t)+Cu(t)-s(t)-b)/\omega(t))\rangle$$
$$+\langle s-s(t), p(t)-(Ax(t)+Cu(t)-s(t)-b)/\omega(t)\rangle\geqq 0, \quad \forall u\in\mathscr{R}^m, \quad \forall s\in[0,\infty)^r,$$

$$p(t+1)=p(t)+(b-Ax(t)-Cu(t)+s(t))/\omega(t),$$

which can be seen as equivalent to

$$0=P^{-1}x(t)-A^Tp(t), \tag{5.7a}$$

$$0=Q^{-1}u(t)-C^Tp(t+1), \tag{5.7b}$$

$$p(t+1)\geqq 0, s(t)\geqq 0, \langle p(t+1), s(t)\rangle=0, \tag{5.7c}$$

$$p(t+1)=p(t)+(b-Ax(t)-Cu(t)+s(t))/\omega(t). \tag{5.7d}$$

Now the conditions (5.7a), (5.7b) are equivalent to, respectively,

$$x(t)=PA^Tp(t), \qquad u(t)=QC^Tp(t+1),$$

which, when combined with (5.7c) and (5.7d), yield

$$(I+CQC^T/\omega(t))p(t+1)-(I-APA^T/\omega(t))p(t)-b/\omega(t)=s(t)/\omega(t)\geqq 0,$$
$$p(t+1)\geqq 0, \langle p(t+1), s(t)\rangle=0.$$

By (5.4a)-(5.4b), the above iteration is exactly (5.3a)-(5.3b). Moreover, Proposition 2(c) shows that $\{p(t)\}$ is well defined and converges to a solution of (5.1). If in addition $J=APA^T$ is positive definite, then Proposition 2(d) shows that $\{p(t)\}$ converges at least linearly. □

Notice that if J is *symmetric* positive semidefinite, then J can always be factored as $J=APA^T$ for some positive definite diagonal matrix P and some matrix A with orthonormal columns (see [21], [47]), so that (5.4a) holds automatically and, in addition, it can be seen that $\sigma=1/\|J\|$ and $\|A\|=1$. If K is *symmetric* positive semidefinite, then by the same argument (5.4b) holds automatically. In this case, the iteration (5.3a)-(5.3b) can be carried out by minimizing a convex quadratic function over the nonnegative orthant. If K is not symmetric but is, say, tridiagonal, then a direct method such as that described in [8] may be used.

To illustrate some of the advantages of the above matrix splitting, suppose that M is of the form

$$M=DPD^T+EQE^T,$$

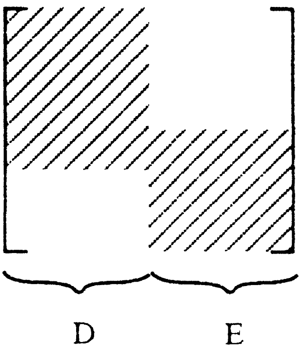


FIG. 1(a). *The matrix* $[D\ \ E]$ *has a staircase structure.*

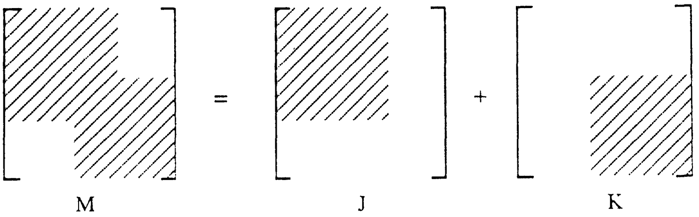


FIG. 1(b). *The matrix M decomposes into the sum of an upper block diagonal matrix J and a lower block diagonal matrix K.*

where P and Q are positive definite diagonal matrices and D and E are matrices of appropriate dimension (such form arises in, for example, quadratic programs with strictly convex separable costs and linear inequality constraints). Suppose that we choose $J=DPD^T$ and $K=EQE^T$. Then, if the matrix $[D\ \ E]$ has the staircase structure shown in Fig. 1(a), the matrices J and K would have, respectively, the upper and lower block diagonal form shown in Fig. 1(b). In this case the problem (5.3a)-(5.3b) is significantly smaller in dimension than the original problem (5.1).

Appendix A. In this appendix we prove Proposition 1. Let p^* be any solution of (2.1). Then, by (2.2),

(A.1a) $$x^*\in\Phi(A'p^*),$$

and there exists some $z^*\in\mathcal{W}$ satisfying

(A.1b) $$z^*\in\Gamma(B'p^*),$$

(A.1c) $$Ax^*+Bz^*=b.$$

From (2.3a)-(2.3c) we also have that, for all t, there holds

(A.2a) $$x(t)\in\Phi(A'p(t)),$$

(A.2b) $$z(t)\in\Gamma(B'p(t+1)).$$

Fix any integer $t\geqq 0$ and, for convenience, let $c=c(t)$. Since $A'p^*\in\Phi^{-1}(x^*)$ and $A'p(t)\in\Phi^{-1}(x(t))$ [cf. (A.1a) and (A.2a)], we have (also using Assumption A(b)) that

$$\begin{aligned}0&=\langle A'p(t)-A'p^*, x(t)-x^*\rangle-\langle A'p(t)-A'p^*, x(t)-x^*\rangle\\&\geqq\sigma\|x(t)-x^*\|^2-\langle p(t)-p^*, Ax(t)-Ax^*\rangle\\&=c\langle Ax(t)-Ax^*, Ax(t)-Ax^*\rangle-\langle p(t)-p^*, Ax(t)-Ax^*\rangle\\&\quad-c\|Ax(t)-Ax^*\|^2+\sigma\|x(t)-x^*\|^2.\end{aligned}$$

Let $\theta=-c\|Ax(t)-Ax^*\|^2+\sigma\|x(t)-x^*\|^2$. The above relation then implies

$$\begin{aligned}0&\geqq\langle -p(t)+c(Ax(t)+Bz(t)-b)+p^*, Ax(t)-Ax^*\rangle\\ &\quad -c\langle Bz(t)-Bz^*, Ax(t)-Ax^*\rangle+\theta\\ &=\langle -\hat{p}(t+1), Ax(t)-Ax^*\rangle-c\langle B\hat{z}(t), A\hat{x}(t)\rangle+\theta,\end{aligned}\tag{A.3}$$

where we let $\hat{x}(t)=x(t)-x^*$, $\hat{z}(t)=z(t)-z^*$, $\hat{p}(t+1)=p(t+1)-p^*$, and the equality follows from (2.3c). Similarly, since $z^*\in\Gamma(B'p^*)$ and $z(t)\in\Gamma(B'p(t+1))$ [cf. (A.1b) and (A.2b)], we have from the monotone property of Γ that

$$\begin{aligned}0&=\langle B'p(t+1)-B'p^*, z(t)-z^*\rangle-\langle B'p(t+1)-B'p^*, z(t)-z^*\rangle\\ &\geqq-\langle p(t+1)-p^*, Bz(t)-Bz^*\rangle\\ &=\langle -\hat{p}(t+1), Bz(t)-Bz^*\rangle.\end{aligned}\tag{A.4}$$

Summing (A.3)-(A.4) and using the fact $Ax^*+Bz^*=b$ (cf. (A.1c)), we obtain

$$\begin{aligned}0&\geqq\langle -\hat{p}(t+1), Ax(t)+Bz(t)-b\rangle-c\langle B\hat{z}(t), A\hat{x}(t)\rangle+\theta\\ &=\langle \hat{p}(t+1), \hat{p}(t+1)-\hat{p}(t)\rangle/c-c\langle B\hat{z}(t), A\hat{x}(t)\rangle+\theta,\end{aligned}$$

where $\hat{p}(t)=p(t)-p^*$ and the equality follows from (2.3c). This, together with the identities (cf. (2.3c))

$$\begin{aligned}2\langle \hat{p}(t+1), \hat{p}(t+1)-\hat{p}(t)\rangle&=\|\hat{p}(t+1)-\hat{p}(t)\|^2+\|\hat{p}(t+1)\|^2-\|\hat{p}(t)\|^2,\\ \|\hat{p}(t+1)-\hat{p}(t)\|^2/c^2&=\|A\hat{x}(t)\|^2+\|B\hat{z}(t)\|^2+2\langle B\hat{z}(t), A\hat{x}(t)\rangle,\end{aligned}$$

implies that

$$0\geqq\|\hat{p}(t+1)\|^2-\|\hat{p}(t)\|^2+c^2\|A\hat{x}(t)\|^2+c^2\|B\hat{z}(t)\|^2+2c\theta,$$

so that, by the definition of c and θ, there holds

$$\begin{aligned}\|\hat{p}(t)\|^2&\geqq\|\hat{p}(t+1)\|^2-c(t)^2\|A\hat{x}(t)\|^2+2c(t)\sigma\|\hat{x}(t)\|^2+c(t)^2\|B\hat{z}(t)\|^2\\ &\geqq\|\hat{p}(t+1)\|^2+c(t)(2\sigma-c(t)\|A\|^2)\|\hat{x}(t)\|^2+c(t)^2\|B\hat{z}(t)\|^2.\end{aligned}$$

Since the choice of t and p^* was arbitrary and, by (2.6), both $2\sigma/\|A\|^2-c(t)$ and $c(t)$ are bounded away from ε, we obtain that, for any solution p^* of (2.1), there holds

$$\|p(t)-p^*\|^2\geqq\|p(t+1)-p^*\|^2+\varepsilon^2\|A\|^2\|x(t)-x^*\|^2+\varepsilon^2\|Bz(t)+Ax^*-b\|^2,\quad\forall t.\tag{A.5}$$

Equation (A.5) shows that $\{p(t)\}$ is bounded and $\sum_{t=0}^{\infty}\|x(t)-x^*\|^2<\infty$, $\sum_{t=0}^{\infty}\|Bz(t)+Ax^*-b\|^2<\infty$, so that

$$\{x(t)\}\to x^*\text{ strongly},\qquad \{Bz(t)\}\to b-Ax^*\text{ strongly}.\tag{A.6}$$

This proves parts (a) and (b).

To prove part (c), note that since (cf. (A.2b))

$$Bz(t)\in B\Gamma(B'p(t+1)),\quad\forall t,$$

and $B\Gamma B'$ is maximal monotone, we have from (A.6) and the limit property for maximal monotone operators (e.g., [5, Prop. 2.5]) that, if p^∞ is any weak limit point of $\{p(t)\}$, then

$$b-Ax^*\in B\Gamma(B'p^\infty).$$

Similarly, we have from (A.2a), (A.6) and the maximal monotone property of $A\Phi A'$ (cf. [19, Prop. 4.1] and the maximal strong monotone property of Φ^{-1}) that

$$Ax^*\in A\Phi(A'p^\infty).$$

Hence, $b \in A\Phi(A'p^\infty)+B\Gamma(B'p^\infty)$ and p^∞ solves (2.1). Then, by using an argument given in [57, proof of Thm. 1] (attributed to Martinet [39]), we obtain that this weak limit point p^∞ is unique. For completeness we give the argument here. Suppose that $\{p(t)\}$ has two weak limit points, say p_1^∞ and p_2^∞. Then, both p_1^∞ and p_2^∞ are solutions of (2.1), so that each can play the role of p^* in (A.5), and we obtain the existence of the limits.

$$\lim_{t\to\infty} \|p(t)-p_k^\infty\| = \mu_k < \infty, \qquad k=1,2.$$

By writing

$$\|p(t)-p_2^\infty\|^2 = \|p(t)-p_2^\infty\|^2 + 2\langle p(t)-p_1^\infty, p_1^\infty - p_2^\infty\rangle + \|p_1^\infty - p_2^\infty\|^2,$$

we see that the limit of $\{\langle p(t)-p_1^\infty, p_1^\infty-p_2^\infty\rangle\}$ must also exist and

$$2\lim_{t\to\infty}\langle p(t)-p_1^\infty, p_1^\infty-p_2^\infty\rangle = (\mu_2)^2-(\mu_1)^2-\|p_1^\infty-p_2^\infty\|^2.$$

Since p_1^∞ is a weak limit point of $\{p(t)\}$, this limit cannot be different from 0. Hence

$$(\mu_2)^2-(\mu_1)^2 = \|p_1^\infty-p_2^\infty\|^2 > 0.$$

By an analogous argument with p_1^∞ and p_2^∞ reversed in role, we also obtain $(\mu_1)^2-(u_2)^2 > 0$. This is a contradiction and hence p^∞ is unique.

Finally we prove part (d). Let p^* be a solution of (2.1). (p^* is unique since $A\Phi A' + B\Gamma B'$ is strongly monotone.) From (A.1a)-(A.1b) and (A.2a)-(A.2b) we have that

$$Ax^* \in A\Phi(A'p^*), \qquad b - Ax^* \in B\Gamma(B'p^*),$$

$$Ax(t) \in A\Phi(A'p(t)), \qquad Bz(t) \in B\Gamma(B'p(t+1)), \quad \forall t.$$

Fix any integer $t \geqq 0$. Since δ and η are the moduli of $A\Phi A'$ and $B\Gamma B'$, respectively, the above relation implies that

$$\langle Ax(t)-Ax^*, p(t)-p^*\rangle \geqq \delta\|p(t)-p^*\|^2,$$

$$\langle Bz(t)+Ax^*-b, p(t+1)-p^*\rangle \geqq \eta\|p(t+1)-p^*\|^2,$$

and hence, by the Cauchy-Schwarz inequality,

$$\|Ax(t)-Ax^*\| \geqq \delta\|p(t)-p^*\|,$$

$$\|Bz(t)+Ax^*-b\| \geqq \eta\|p(t+1)-p^*\|.$$

This together with (A.5) yields

$$\|p(t)-p^*\|^2 \geqq \|p(t+1)-p^*\|^2 + \varepsilon^2\delta^2\|p(t)-p^*\|^2 + \varepsilon^2\eta^2\|p(t+1)-p^*\|^2,$$

so that (since either δ or η is positive) $\|p(t)-p^*\|^2$ converges to zero at least linearly with a convergence ratio of $(1-\varepsilon^2\delta^2)/(1+\varepsilon^2\eta^2)$. Since we also have from (A.5) that

$$\|p(t)-p^*\|^2 \geqq \varepsilon^2\|A\|^2\|x(t)-x^*\|^2 + \varepsilon^2\|Bz(t)+Ax^*-b\|^2,$$

so that both $\|x(t)-x^*\|$ and $\|Bz(t)+Ax^*-b\|$ are upper bounded by some constant times $\|p(t)-p^*\|$, then part (d) follows. (Note that, in this case, $\{p(t)\}$ converges in the strong topology.) □

Acknowledgments. Thanks are due to D. P. Bertsekas and J. Eckstein for their many insightful comments. I am particularly indebted to the former for introducing me to this general area. Thanks are also due to an anonymous referee whose comments led to many improvements in the paper.

REFERENCES

[1] A. Auslender, *Optimisation: Méthodes Numériques*, Masson, Paris, 1976.

[2] M. L. Balinski and R. W. Cottle, Eds., *Complementarity and Fixed Point Problems*, Mathematical Programming Study 7. North-Holland, Amsterdam, 1978.

[3] D. P. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

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

[5] H. Brézis, *Operateurs Maximaux Monotones*, North-Holland, Amsterdam, 1973.

[6] R. E. Bruck, Jr., *An iterative solution of a variational inequality for certain monotone operators in Hilbert space*, Bull. Amer. Math. Soc., 81 (1975), pp. 890-892.

[7] ———, *On the weak convergence of an ergodic iteration for the solution of variational inequalities for monotone operators in Hilbert space*, J. Math. Anal. Appl., 61 (1977), pp. 159-164.

[8] R. W. Cottle and R. S. Sacher, *On the solution of large, structured, linear complementarity problems: the tridiagonal case*, J. Appl. Math. Optim., 3 (1977), pp. 321-340.

[9] R. W. Cottle, F. Giannessi, and J.-L. Lions, Eds., *Variational Inequalities and Complementarity Problems: Theory and Applications*, Wiley, New York, 1980.

[10] S. Dafermos, *An iterative scheme for variational inequalities*, Math. Programming, 26 (1983), pp. 40-47.

[11] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, New York, Heidelberg, Tokyo, 1985.

[12] V. Dolezal, *Monotone Operators and Applications in Control and Network Theory*, Elsevier Scientific, Amsterdam, 1979.

[13] J. Eckstein and D. P. Bertsekas, *On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators*, Report P-1919, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, October, 1989.

[14] J. Eckstein, *Splitting methods for monotone operators with applications to parallel optimization*, Ph.D. Thesis, Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, June, 1989.

[15] Everett, *Generalized Lagrange multiplier method for solving problems of optimum allocation of resources*, Oper. Res., 11 (1963), pp. 399-417.

[16] M. Fortin and R. Glowinski, Eds., *Augmented Lagrangian Methods: Applications to the Solution of Boundary-Valued Problems*, North-Holland, Amsterdam, 1983.

[17] D. Gabay, *Méthodes numériques pour l'optimisation non linéaire*, Thèse de Doctorat d'Etat et Science Mathématiques, Université Pierre et Marie Curie, Paris VI, 1979.

[18] ———, *Methods for the decomposition of variational inequalities via the proximal point algorithm*, Abstract, 10th International Symposium on Mathematical Programming, Montreal, 1979.

[19] ———, *Applications of the method of multipliers to variational inequalities*, in Augmented Lagrangian Methods: Applications to the Solution of Boundary-Valued Problems, M. Fortin and R. Glowinski, eds., North-Holland, Amsterdam, 1983, pp. 299-331.

[20] D. Gabay and B. Mercier, *A dual algorithm for the solution of nonlinear variational problems via finite-element approximations*, Comp. Math. Appl., 2 (1976), pp. 17-40.

[21] F. R. Gantmacher, *The Theory of Matrices*, Vols. 1-2, Chelsea, New York, 1959.

[22] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, New York, Berlin, Heidelberg, Tokyo, 1984.

[23] R. Glowinski and P. Le Tallec, *Augmented Lagrangian methods for the solution of variational problems*, Mathematics Research Center Technical Summary Report #2965, University of Wisconsin, Madison, WI, January 1987.

[24] R. Glowinski, P. L. Lions, and R. Tremolieres, *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.

[25] R. GLOWINSKI AND A. MARROCCO, *Sur l'approximation par éléments finis d'ordre un, et la résolution par pénalisation-dualité d'une classe de problèmes de Dirichlet nonlinéaires*, Rev. Française d'Aut. Inf. Rech. Opér., R-2 (1975), pp. 41-76.

[26] A. A. GOLDSTEIN, *Convex programming in Hilbert spaces*, Bull. Amer. Math. Soc., 70 (1964), pp. 709-710.

[27] P. C. HAARHOFF AND J. D. BUYS, *A new method for the optimization of a nonlinear function subject to nonlinear constraints*, The Computer Journal, 13 (1970), pp. 178-184.

[28] S. P. HAN AND G. LOU, *A parallel algorithm for a class of convex programs*, SIAM J. Control Optim., 26 (1988), pp. 345-355.

[29] M. R. HESTENES, *Multiplier and Gradient Methods*, J. Optim. Theory Appl., 4 (1969), pp. 303-320.

[30] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[31] L. S. LASDON, *Optimization Theory for Large Systems*, Macmillan, New York, 1970.

[32] B. LEMAIRE, *Coupling optimization methods and variational convergence*, in Trends in Mathematical Optimization, K.-H. Hoffman, J.-B. Hiriart-Urruty, J. Zowe, C. Lemarechal, eds., Birkhäuser Verlag, Basel, 1988, pp. 163-179.

[33] Y. Y. LIN AND J.-S. PANG, *Iterative methods for large convex quadratic programs: A survey*, SIAM J. Control Optim., 25 (1987), pp. 383-411.

[34] P. L. LIONS AND B. MERCIER, *Splitting algorithms for the sum of two nonlinear operators*, SIAM J. Numer. Anal., 16 (1979), pp. 964-979.

[35] P. L. LIONS AND R. TEMAM, *Une méthode d'éclatement des opérateurs et des contraintes en calcul des variations*, C.R. Acad. Sci. Paris, 263 (1966), pp. 563-565.

[36] D. G. LUENBERGER, *Introduction to Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1984.

[37] O. L. MANGASARIAN, *Solution of symmetric linear complementarity problems by iterative methods*, J. Optim. Theory Appl., 22 (1977), pp. 465-485.

[38] O. L. MANGASARIAN AND R. DE LEONE, *Serial and parallel solution of large scale linear programs by augmented lagrangian successive overrelaxation*, Computer Sciences Technical Report 701, University of Wisconsin, Madison, WI, June 1987.

[39] B. MARTINET, *Regularisation d'inéquations variationelles par approximations successives*, Rev. Française d'Aut. Inf. Rech. Opér., (1970), pp. 154-159.

[40] ———, *Détermination approchée d'un point fixe d'une application pseudo-contractante*, C.R. Acad. Sci. Paris, 274 (1972), pp. 163-165.

[41] G. J. MINTY, *Monotone (nonlinear) operators in Hilbert space*, Duke Math. J., 29 (1962), pp. 341-346.

[42] ———, *On the monotonicity of the gradient of a convex function*, Pacific J. Math., 14 (1964), pp. 243-247.

[43] J. J. MOREAU, *Proximité et dualité dans un espace Hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273-299.

[44] K. MOUALLIF, V. H. NGUYEN, AND J.-J. STRODIOT, *A perturbed parallel decomposition method for a class of nonsmooth convex minimization problems*, Report 88/25, Département de Mathématique, Facultés Universitaires de Namur, Namur, Belgium, December 1988.

[45] K. G. MURTY, *Linear Complementarity, Linear and Nonlinear Programming*, Helderman-Verlag, Berlin, 1988.

[46] Z. OPIAL, *Weak convergence of the sequence of successive approximations for nonexpansive mappings*, Bull. Amer. Math. Soc., 73 (1967), pp. 591-597.

[47] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[48] J.-S. PANG, *On the convergence of a basic iterative method for the implicit complementarity problem*, J. Optim. Theory Appl., 37 (1982), pp. 149-162.

[49] J.-S. PANG AND D. CHAN, *Iterative methods for variational and complementarity problems*, Math. Prog., 24 (1982), pp. 284-313.

[50] D. PASCALI AND S. SBURLAN, *Nonlinear Mappings of Monotone Type*, Editura Academeie, Bucharest, 1978.

[51] G. B. PASSTY, *Ergodic convergence to a zero of the sum of monotone operators in Hilbert space*, J. Math. Anal. Appl., 72 (1979), pp. 383-390.

[52] B. T. POLJAK, *Introduction to Optimization*, Optimization Software Inc., New York, 1987.

[53] M. J. D. POWELL, *A method for nonlinar constraints in minimization problems*, in Optimization, R. Fletcher, ed., Academic Press, NY, 1969, pp. 283-298.

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

[55] ———, *On the maximality of sums of nonlinear monotone operators*, Trans. Amer. Math. Soc. 149 (1970), pp. 75-88.

[56] ———, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97–116.

[57] ———, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877–898.

[58] J. E. SPINGARN, *Applications of the method of partial inverses to convex programming: Decomposition*, Math. Prog., 32 (1985), pp. 199–223.

[59] R. TEMAM, *Sur la stabilité et la convergence de la méthode des pas fractionnaires*, Annali di Matematica Pura ed Applicata, LXXIX (1968), pp. 191–380.

[60] P. TSENG, *Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming*, Report P-1866, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, June 1989.

[61] H. UZAWA, *Iterative methods for concave programming*, in Studies in Linear and Nonlinear Programming, K. J. Arrow, L. Hurwicz, and H. Uzawa, eds., Stanford University Press, Stanford, CA, 1958, pp. 154–165.

[62] N. Young, *An Introduction to Hilbert Space*, Cambridge University Press, Cambridge, 1988.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 1, pp. 139–152, January 1991

007

ON THE LINEAR QUADRATIC GAUSSIAN PROBLEM WITH CORRELATED NOISE AND ITS RELATION TO MINIMUM VARIANCE CONTROL*

R. H. KWONG†

Abstract. The linear quadratic Gaussian (LQG) stochastic control problem with correlated dynamic and observation noise and no information delay is studied. An explicit feedback solution is given for finite as well as infinite time problems. These results are then applied to minimum variance control of single-input single-output ARMAX systems. The LQG controller and the minimum variance controller obtained using input–output methods are shown to be identical for any system delay, extending a result of [5].

Key words. LQG stochastic control, correlated noise, minimum variance control, ARMAX systems, certainty equivalence

AMS(MOS) subject classifications. 93E11, 93E20

1. Introduction. The linear quadratic Gaussian (LQG) problem has been extensively discussed in the literature [2], [4], [6], [7], [14]. The standard formulation of the problem assumes either that the dynamic and the observation noise processes are uncorrelated, or that the control at time k is a function of the observations up to time $k-1$. There does not seem to be an explicit solution published for the problem in which the dynamic and observation noise processes were correlated, and that the control at time k is allowed to depend on observations up to time k. This problem we shall henceforth refer to as the general LQG problem. The only result previously published known to this author is given in [4]. However, this reference treats the special case in which the quadratic cost does not have any terms involving the state, except for the terminal penalty. Furthermore, the system matrix is assumed to be invertible, and the solution given is not explicitly in feedback form. The method of introducing output feedback and a new dynamic noise process uncorrelated with the observation noise is mentioned in [14]. However, no explicit solution is given.

The presence of this gap in the LQG theory is rather surprising in that one of the standard problems in stochastic control, minimum variance control for ARMAX systems, when the ARMAX system is described in the innovations form [6], [7], is a problem of this type. There exists a large body of literature on minimum variance control of ARMAX systems [2], [8], [9], [13], [17]. Many results for multivariable systems, minimum and nonminimum phase systems, as well as systems with general cost weightings have been developed. These results have been further extended to treat self-tuning control problems [3], [10]. Many successful applications of this control scheme have also been documented [2], [11]. However, most of these treatments use an input–output polynomial matrix approach. The connection between the LQG problem formulated in state space form and that formulated in input–output form does not seem to have been a topic of specific concern in these papers, and has not been explicitly considered. In fact, there appears to have been some doubt expressed in the literature [19] as to whether or not linear regulator design based on the innovations state space model for ARMAX systems will lead to the same optimal control law as that obtained using an input–output polynomial approach. In [5], such an explicit

* Received by the editors February 27, 1989; accepted for publication February 7, 1990.

† Department of Electrical Engineering, University of Toronto, Toronto, Ontario, Canada M5S 1A4.
This research was supported by Natural Sciences and Engineering Research Council of Canada grant A0875.

connection between polynomial and state space formulations of minimum variance control was explored. However, that result was not obtained using a control gain derived from the solution of a Riccati equation, so that strictly speaking, it did not fully map the connections between state space LQG theory and polynomial minimum variance control. Moreover, the treatment is restricted to only the unit delay case.

In this paper, we solve the general LQG problem and explore extensions and applications of these results. In § 2, we formulate the stochastic control problem and point out the complications introduced by the correlation between the dynamic and observation noise. In § 3, we solve the stochastic control problem over a finite time interval, giving the optimal control law in feedback form. The result is given a certainty equivalence interpretation using an augmented state. The structure of the optimal control allows for easy extension to infinite time problems. In particular, the steady state version of the finite time control law is shown to be optimal for the average cost per unit time problem. These results are then applied to the minimum variance control problem for scalar ARMAX systems in § 4. We prove the "folk theorem" that the state space LQG theory gives the identical control law as the usual polynomial approach, for any delay $d \geqq 1$. As far as we know, this is the first complete proof of this folk theorem in the general delay case, extending the results of [5]. Section 5 contains the concluding remarks.

2. The general LQG problem. We begin by stating some well-known results from LQG theory, [6], [7], mainly to establish notation.

Consider the linear stochastic system

$$x_{k+1} = A_k x_k + B_k u_k + C_k w_k, \qquad x_{k_0} = x_0 \tag{2.1a}$$

$$y_k = H_k x_k + e_k; \tag{2.1b}$$

w_k is assumed to be zero mean normalized Gaussian white noise, i.e., $E w_k w_j^T = I\delta_{kj}$, where δ_{kj} is the Kronecker delta function: $\delta_{kj} = 1$ if $k = j$, 0 otherwise. e_k is assumed to be zero mean Gaussian white noise with $E e_k e_j^T = R_k \delta_{kj}$. w_k and e_k are allowed to be correlated with $E e_k w_j^T = G_k \delta_{kj}$. The initial state x_0 is a Gaussian random variable independent of w_k and e_k, with mean m_0 and covariance Σ_0. R_k is assumed to satisfy $R_k \geqq \delta I$, some $\delta > 0$. The control u_k is allowed to take the form $u_k = \phi_k(y^k)$ where $y^k = \{y_{k_0}, y_{k_0+1}, \cdots, y_k\}$. The objective of control is to find, in the class of admissible control laws, the one that minimizes

$$J = E\left\{ x_N^T Q x_N + \sum_{k=k_0}^{N-1} \|D_k x_k + F_k u_k\|^2 \right\} \tag{2.2}$$

where $Q \geqq 0$ and $\|x\|$ denotes the Euclidean vector norm $(x^T x)^{1/2}$. It is well-known [6], [7], that the conditional means $\hat{x}_{k/k-1} = E(x_k/y^{k-1})$ and $\hat{x}_{k/k} = E(x_k/y^k)$ are generated by the following Kalman filtering equations

$$\hat{x}_{k/k} = \hat{x}_{k/k-1} + P_k H_k^T (H_k P_k H_k^T + R_k)^{-1} \nu_k \tag{2.3a}$$

$$\hat{x}_{k+1/k} = A_k \hat{x}_{k/k} + B_k u_k + C_k G_k^T (H_k P_k H_k^T + R_k)^{-1} \nu_k \tag{2.3b}$$

$$\hat{x}_{k_0/k_0-1} = m_0 \tag{2.3c}$$

$$\nu_k = y_k - H_k \hat{x}_{k/k-1}. \tag{2.3d}$$

The innovations process ν_k is known to be a white noise process [6], [7]. For notational convenience, we write

$$\Lambda_k = H_k P_k H_k^T + R_k \tag{2.4}$$

and

$$K_k=(A_kP_kH_k^T+C_kG_k^T)\Lambda_k^{-1}. \tag{2.5}$$

The error covariance $P_k=E[(x_k-\hat{x}_{k/k-1})(x_k-\hat{x}_{k/k-1})^T]$ satisfies the filtering Riccati difference equation

$$P_{k+1}=A_kP_kA_k^T+C_kC_k^T-K_k\Lambda_kK_k^T \tag{2.6a}$$

$$P_{k_0}=\Sigma_0. \tag{2.6b}$$

It is also well known [6], [7] that if admissible control laws are restricted to be of the form $u_k=\phi_k(y^{k-1})$, so that there is a one-step information delay, the optimal control law is given by

$$u_k=-(B_k^TS_{k+1}B_k+F_k^TF_k)^{-1}(B_k^TS_{k+1}A_k+F_k^TD_k)\hat{x}_{k/k-1} \tag{2.7}$$

where S_k satisfies the control Riccati difference equation

$$S_k=A_k^TS_{k+1}A_k+D_k^TD_k-L_k^T\Gamma_kL_k \tag{2.8a}$$

$$S_N=Q \tag{2.8b}$$

with

$$\Gamma_k=B_k^TS_{k+1}B_k+F_k^TF_k \tag{2.9}$$

assumed to be invertible for all k (guaranteed, for example, if $F_k^TF_k\geqq\delta I$ for some $\delta>0$) and

$$L_k=\Gamma_k^{-1}(B_k^TS_{k+1}A_k+F_k^TD_k). \tag{2.10}$$

The above result does not require any conditions on the product $C_kG_k^T$ and is one standard formulation of the certainty equivalence principle. On the other hand, if u_k is allowed to be of the form $\phi_k(y^k)$ but with $C_kG_k^T=0$, then the optimal control law is given by

$$u_k=-L_k\hat{x}_{k/k}. \tag{2.11}$$

This is the standard formulation of the certainty equivalence principle for the case of no information delay [4], [14].

If the system matrices are constant, and the criterion is

$$J_{av}=\lim_{N\to\infty}\frac{1}{N}E\sum_{k=0}^{N-1}\|Dx_k+Fu_k\|^2 \tag{2.12}$$

under the appropriate structural assumptions of stabilizability and detectability, the steady-state versions of (2.7) and (2.11) are again optimal for the one-step observation delay and no delay cases, respectively, [6], [7], [14].

However, in the general case, where $C_kG_k^T\neq 0$ and $u_k=\phi_k(y^k)$, the control law (2.11) will not be optimal. This is illustrated by the following simple example, using the criterion J_{av}.

$$x_{k+1}=u_k+cw_k \qquad |c|<1$$

$$y_k=x_k+w_k.$$

The criterion is

$$J_{av}=\lim_{N\to\infty}\frac{1}{N}E\sum_{k=0}^{N-1}x_k^2.$$

For the standard form of certainty equivalence control (2.11), the steady-state version yields $u_k = 0$, resulting in $J_{av} = c^2$. However, the optimal control law is in fact $u_k = -cy_k$, resulting in $x_{k+1} = -cx_k$ and $J_{av} = 0$.

Let us look at the standard approach [6], [7], [14] to derive the optimal control law in the partial observations case. It involves reformulating the partial observations problem into a perfect observations problem using the optimal estimate as the new state. Since there is no information delay, we expect that $\hat{x}_{k/k}$ should be involved. If we write a state equation for $\hat{x}_{k/k}$, from (2.3a, (2.3b) we get

$$\hat{x}_{k+1/k+1} = A_k\hat{x}_{k/k} + B_k u_k + C_k G_k^T \Lambda_k^{-1} \nu_k + P_{k+1} H_{k+1}^T \Lambda_{k+1}^{-1} \nu_{k+1}. \tag{2.13}$$

The noise process on the right-hand side of (2.13) is, however, not white. This is why the standard equivalence control law is not directly applicable here.

3. Explicit feedback solution of the general LQG problem. Even though in (2.13) we do not get a standard state space model with white noise input for the process $\hat{x}_{k/k}$, it does suggest the approach to solving the problem. Consider the augmented state $\xi_k = [\hat{x}_{k/k}^T \quad \nu_k^T]^T$. The process ξ_k satisfies the equation

$$\xi_{k+1} = \begin{bmatrix} A_k & C_k G_k^T \Lambda_k^{-1} \\ 0 & 0 \end{bmatrix} \xi_k + \begin{bmatrix} B_k \\ 0 \end{bmatrix} u_k + \begin{bmatrix} P_{k+1} H_{k+1}^T \Lambda_{k+1}^{-1} \\ I \end{bmatrix} \nu_{k+1}. \tag{3.1}$$

If we now define the matrices

$$\bar{A}_k = \begin{bmatrix} A_k & C_k G_k^T \Lambda_k^{-1} \\ 0 & 0 \end{bmatrix}$$

$$\bar{B}_k = \begin{bmatrix} B_k \\ 0 \end{bmatrix}$$

$$\bar{C}_k = \begin{bmatrix} P_{k+1} H_{k+1}^T \Lambda_{k+1}^{-1} \\ I \end{bmatrix}$$

then ξ_k satisfies the standard state equation with white noise input ν_{k+1}

$$\xi_{k+1} = \bar{A}_k \xi_k + \bar{B}_k u_k + \bar{C}_k \nu_{k+1}. \tag{3.2}$$

Observe that ξ_k is known given y^k.

We can now follow [6], [7], [14] to transform the quadratic cost criterion in terms of the observed state process ξ_k and u_k as follows:

$$J = E\left[\hat{x}_{N/N}^T Q \hat{x}_{N/N} + \sum_{k=k_0}^{N-1} \|D_k \hat{x}_{k/k} + F_k u_k\|^2\right] + \operatorname{tr}\left(Q P_{N/N} + \sum_{k=k_0}^{N-1} D_k P_{k/k} D_k^T\right) \tag{3.3}$$

where $P_{k/k} = E(x_k - \hat{x}_{k/k})(x_k - \hat{x}_{k/k})^T$ is related to P_k by

$$P_{k/k} = P_k - P_k H_k^T \Lambda_k^{-1} H_k P_k. \tag{3.4}$$

Note that the last two terms of (3.3) are independent of u_k. Hence by defining $\bar{D}_k = [D_k \quad 0]$, $\bar{Q} = \left[\begin{smallmatrix} Q & 0 \\ 0 & 0 \end{smallmatrix}\right]$, we see that minimizing J is equivalent to minimizing $\bar{J}$

$$\bar{J} = E\left[\xi_N^T \bar{Q} \xi_N + \sum_{k=k_0}^{N-1} \|\bar{D}_k \xi_k + F_k u_k\|^2\right]. \tag{3.5}$$

The following result, solving the general LQG problem over a finite time interval, may now be stated.

THEOREM 1. *The optimal control law for the general* LQG *problem in the class of control laws of the form* $u_k = \phi_k(y^k)$ *is given by*

$$u_k = -(B_k^T S_{k+1} B_k + F_k^T F_k)^{-1}[(B_k^T S_{k+1} A_k + F_k^T D_k)\hat{x}_{k/k} + B_k^T S_{k+1} C_k G_k^T (H_k P_k H_k^T + R_k)^{-1} \nu_k]. \tag{3.6}$$

Equivalently, the optimal control law can also be written as

$$u_k = -L_k \hat{x}_{k/k-1} - (L_k P_k H_k^T + \Gamma_k^{-1} B_k^T S_{k+1} C_k G_k^T)\Lambda_k^{-1} \nu_k \tag{3.7a}$$

or

$$u_k = -L_k \hat{x}_{k/k} - \Gamma_k^{-1} B_k^T S_{k+1} C_k G_k^T R_k^{-1}(y_k - H\hat{x}_{k/k}) \tag{3.7b}$$

where Γ_k *and* L_k *are defined in* (2.9) *and* (2.10), *respectively.*

Proof. It follows from (3.2) and (3.5) and the standard LQG theory for perfect state observations [6], [7], [14] that the optimal control law is given by

$$u_k = -(\bar{B}_k^T \bar{S}_{k+1} \bar{B}_k + F_k^T F_k)^{-1}[\bar{B}_k^T \bar{S}_{k+1} \bar{A}_k + F_k^T \bar{D}_k]\xi_k \tag{3.8}$$

where $\bar{S}_k$ satisfies the Riccati difference equation analogous to (2.8) with $\bar{A}_k$, $\bar{B}_k$, $\bar{D}_k$, $\bar{Q}$ replacing A_k, B_k, D_k, and Q, respectively. Partitioning $\bar{S}_k$ into

$$\bar{S}_k = \begin{bmatrix} S_1(k) & S_2(k) \\ S_2^T(k) & S_3(k) \end{bmatrix}$$

it is readily seen, by a straightforward calculation, that $S_1(k)$ is in fact just S_k of (2.8). Since $\bar{B}_k^T \bar{S}_{k+1} \bar{A}_k = [B_k^T S_1(k+1) A_k \quad B_k^T S_1(k+1) C_k G_k^T \Lambda_k^{-1}]$ and $\bar{B}_k^T \bar{S}_{k+1} \bar{B}_k + F_k^T F_k = \Gamma_k$, substitution into (3.8) gives (3.6). Using (2.3a) and the definitions of L_k, P_k, and K_k, (3.6) can immediately be rewritten as (3.7a). To show (3.7b), we note the easily verified relation

$$(H_k P_k H_k^T + R_k)^{-1} \nu_k = R_k^{-1}(y_k - H_k \hat{x}_{k/k}).$$

On substituting this equation into (3.6), (3.7b) follows.

Observe that (3.6) reduces to the standard certainty equivalence control law (2.11) if $C_k G_k^T = 0$. Although (2.11) is not optimal for the general LQG problem, we will now give an interpretation of (3.7b) as a certainty equivalence control law for an augmented system.

Since we know [6], [7], [14] that the standard certainty equivalence principle applies when there is no correlation between dynamic and observation noise, we first transform dynamic noise so that it becomes uncorrelated with the observation noise.

Let $\varepsilon_k = w_k - G_k^T R_k^{-1} e_k$. Then the system equation (2.1) can be written in the form

$$x_{k+1} = A_k x_k + B_k u_k + C_k G_k^T R_k^{-1}(y_k - H_k x_k) + C_k \varepsilon_k. \tag{3.9}$$

The process ε_k is uncorrelated with e_k. We now ask the following question. The results of Theorem 1 give the optimal control as a feedback law on a state estimate. Can we find an appropriate state and its corresponding perfectly observed linear quadratic optimal control problem, whose solution gives the *same* feedback law as Theorem 1, but now on the perfectly observed state? The choice is suggested by the optimal feedback law of (3.7b). Define the augmented state

$$p_k = [x_k^T \quad y_k^T]^T.$$

The right-hand side of (3.7b) can be interpreted as a feedback law on the optimal estimate of p_k since $E(p_k/y^k) = [\hat{x}_{k/k}^T \quad y_k^T]^T$. This suggests that p_k may be the appropriate state process.

We can write the following equation for p_k.

$$\begin{bmatrix} I & 0 \\ -H_{k+1} & I \end{bmatrix}\begin{bmatrix} x_{k+1} \\ y_{k+1} \end{bmatrix} = \begin{bmatrix} A_k - C_kG_k^TR_k^{-1}H_k & C_kG_k^TR_k^{-1} \\ 0 & 0 \end{bmatrix}\begin{bmatrix} x_k \\ y_k \end{bmatrix} + \begin{bmatrix} B_k \\ 0 \end{bmatrix} w_k + \begin{bmatrix} C_k\varepsilon_k \\ e_{k+1} \end{bmatrix}. \tag{3.10}$$

This is in the form

$$E_{k+1}p_{k+1} = \tilde{A}_kp_k + \tilde{B}_ku_k + \theta_k \tag{3.11}$$

with obvious definitions for E_{k+1}, $\tilde{A}_k$, $\tilde{B}_k$, and θ_k. Since E_k^{-1} exists for all k, we can also write (3.11) in state form as follows:

$$p_{k+1} = E_{k+1}^{-1}\tilde{A}_kp_k + E_{k+1}^{-1}\tilde{B}_ku_k + E_{k+1}^{-1}\theta_k. \tag{3.12}$$

The above discussion suggests that we solve the optimal control problem, assuming that p_k is perfectly observed, and connect the resulting feedback law to that given in (3.7b). Define $\tilde{D}_k = [D_k \quad 0]$ and $\tilde{Q} = [\begin{smallmatrix} Q & 0 \\ 0 & 0 \end{smallmatrix}]$. The cost criterion for the perfectly observed problem becomes

$$J_p = E\left\{\sum_{k=0}^{N-1} \|\tilde{D}_kp_k + F_ku_k\|^2 + p_N^T\tilde{Q}p_N\right\}.$$

The associated control Riccati equation is given by

$$\begin{aligned} \tilde{S}_k = {} & \tilde{A}_k^TE_{k+1}^{-T}\tilde{S}_{k+1}E_{k+1}^{-1}\tilde{A}_k - (\tilde{A}_k^TE_{k+1}^{-T}\tilde{S}_{k+1}E_{k+1}^{-1}\tilde{B}_k + \tilde{D}_k^TF_k) \\ & \cdot (\tilde{B}_k^TE_{k+1}^{-T}\tilde{S}_{k+1}E_{k+1}^{-1}\tilde{B}_k + F_k^TF_k)^{-1} \\ & \cdot (\tilde{B}_k^TE_{k+1}^{-T}\tilde{S}_{k+1}E_{k+1}^{-1}\tilde{A}_k + F_k^T\tilde{D}_k) + \tilde{D}_k^T\tilde{D}_k \end{aligned} \tag{3.13}$$

where $E_{k+1}^{-T} = (E_{k+1}^{-1})^T$.

Let

$$\hat{S}_k = E_k^{-T}\tilde{S}_kE_k^{-1}.$$

We will need the following lemma.

LEMMA 1. *Partition*

$$\hat{S}_k = \begin{bmatrix} \hat{S}_k^{11} & \hat{S}_k^{12} \\ \hat{S}_k^{21} & \hat{S}_k^{22} \end{bmatrix}.$$

Then $\hat{S}_k^{11} = S_k$ *where* S_k *is the solution to the original Riccati equation* (2.8).

Proof. First we observe that

$$E_k^{-1} = \begin{bmatrix} I & 0 \\ H_k & I \end{bmatrix}.$$

Hence

$$\tilde{A}_kE_k^{-1} = \begin{bmatrix} A_k & C_kG_k^TR_k^{-1} \\ 0 & 0 \end{bmatrix} = \hat{A}_k, \qquad \tilde{D}_kE_k^{-1} = \tilde{D}_k.$$

On premultiplying by E_k^{-T} and post-multiplying by E_k^{-1} in (3.13), we have

$$\begin{aligned} \hat{S}_k = {} & \hat{A}_k^T\hat{S}_{k+1}\hat{A}_k - (\hat{A}_k^T\hat{S}_{k+1}\tilde{B}_k + \tilde{D}_k^TF_k)(\tilde{B}_k^T\hat{S}_{k+1}\tilde{B}_k + F_k^TF_k)^{-1} \\ & \cdot (\tilde{B}_k^T\hat{S}_{k+1}\hat{A}_k + F_k^T\tilde{D}_k) + \tilde{D}_k^T\tilde{D}_k^T. \end{aligned} \tag{3.14}$$

Using the structure of $\hat{A}_k$, $\tilde{B}_k$, and $\tilde{D}_k$, we find that

$$\tilde{B}_k^T \hat{S}_{k+1} \tilde{B}_k + F_k^T F_k = B_k^T \hat{S}_{k+1}^{11} B_k + F_k^T F_k \tag{3.15}$$

$$\tilde{B}_k^T \hat{S}_{k+1} \hat{A}_k + F_k^T \tilde{D}_k = A_k^T \hat{S}_{k+1}^{11} B_k + F_k^T D_k. \tag{3.16}$$

On substitution into (3.14), we find $\hat{S}_k^{11}$ satisfies exactly the same equation as S_k. The lemma is proved.

We can now establish the precise connection between the solution of the perfectly observed problem with state p_k and cost criterion J_p, and the optimal control law (3.7b) of Theorem 1. The optimal control law for the system (3.12) with cost criterion J_p is given by

$$\begin{aligned} u_k &= -(\tilde{B}_k^T E_{k+1}^{-T} \bar{S}_{k+1} E_{k+1}^{-1} \tilde{B}_k + F_k^T F_k)^{-1} (\tilde{B}_k^T E_{k+1}^{-T} \bar{S}_{k+1} E_{k+1}^{-1} \tilde{A}_k + F_k^T \tilde{D}_k) p_k \\ &= -(\tilde{B}_k^T \hat{S}_{k+1} \tilde{B}_k + F_k^T F_k)^{-1} (\tilde{B}_k^T \hat{S}_{k+1} \tilde{A}_k + F_k^T \tilde{D}_k) p_k. \end{aligned} \tag{3.17}$$

Using Lemma 1 in (3.17), we find that

$$\begin{aligned} u_k &= -\Gamma_k^{-1} \{ B_k^T S_{k+1} [(A_k - C_k G_k^T R_k^{-1} H_k) x_k + C_k G_k^T R_k^{-1} y_k] + F_k^T D_k x_k \} \\ &= -L_k x_k - \Gamma_k^{-1} B_k^T S_{k+1} C_k G_k^T R_k^{-1} (y_k - H x_k). \end{aligned} \tag{3.18}$$

On comparing (3.18) with (3.7b), we see that two feedback laws are identical. Thus the optimal control law (3.7b) for the general LQG problem can be interpreted as a certainty equivalence control law with the augmented state $p_k = [x_k^T \quad y_k^T]^T$ in (3.17) replaced by $\hat{p}_{k/k} = [\hat{x}_{k/k}^T \quad y_k^T]^T$.

Next we turn to infinite time problems. The one we will be primarily interested in is the average cost per unit time problem, although results for the discounted cost problem can be obtained as well using similar techniques. We assume the system to be time-invariant so that it is described by

$$x_{k+1} = A x_k + B u_k + C w_k \tag{3.19a}$$

$$y_k = H x_k + e_k \tag{3.19b}$$

with the cost criterion J_{av} given by (2.12). Admissible control laws u_k are of the form $\phi_k(y^k)$ such that J_{av} exists and that $\lim_{N\to\infty} (1/N) E\|x_N\|^2 = 0$. For the rest of this section, we assume $F^T F > 0$.

Define

$$\begin{aligned} \hat{A} &= A - B(F^T F)^{-1} F^T D \\ \hat{D} &= D - F(F^T F)^{-1} F^T D \\ \check{A} &= A - C G^T R^{-1} H \\ \check{C} &= C[I - G^T R^{-1} G]^{1/2}. \end{aligned}$$

It is known [7] that if (A, B) is stabilizable and $(\hat{D}, \hat{A})$ detectable, there exists a unique positive semidefinite solution to the control algebraic Riccati equation (CARE)

$$S = A^T S A + D^T D - L^T \Gamma L \tag{3.20}$$

where

$$\begin{aligned} \Gamma &= B^T S B + F^T F \\ L &= \Gamma^{-1} (B^T S A + F^T D). \end{aligned}$$

Similarly, if $(\check{A}, \check{C})$ is stabilizable and (H, A) detectable, there exists a unique positive semidefinite solution to the filter algebraic Riccati equation (FARE)

$$P = APA^T + CC^T - K\Lambda K^T \tag{3.21}$$

where

$$\Lambda = HPH^T + R$$

and

$$K = (APH^T + CG^T)\Lambda^{-1}.$$

Let $M_k = (LP_kH^T + \Gamma^{-1}B^TSCG^T)\Lambda_k^{-1}$ and $M = (LPH^T + \Gamma^{-1}B^TSCG^T)\Lambda^{-1}$. Then it is also true that $\lim_{k\to\infty} K_k = K$ and $\lim_{k\to\infty} M_k = M$.

As one may expect, the solution to the average cost per unit time problem is given by the steady-state version of (3.6) or (3.7).

THEOREM 2. *Assume that (A, B) and $(\check{A}, \check{C})$ are stabilizable, and that (H, A) and $(\hat{D}, \hat{A})$ are detectable. Then the optimal control law for the average cost per unit time problem is given by*

$$\begin{aligned} u_k &= -(L\hat{x}^s_{k/k-1} + M\bar{\nu}_k) \\ &= -(L\hat{x}^s_{k/k} + \Gamma^{-1}B^TSCG^T\Lambda^{-1}\bar{\nu}_k) \\ &= -[L\hat{x}^s_{k/k} + \Gamma^{-1}B^TSCG^TR^{-1}(y_k - H\hat{x}^s_{k/k})] \end{aligned} \tag{3.22}$$

where $\hat{x}^s_{k/k-1}$, $\hat{x}^s_{k/k}$, and $\bar{\nu}_k$ are generated by the steady-state Kalman filter

$$\hat{x}^s_{k+1/k} = A\hat{x}^s_{k/k-1} + Bu_k + K\bar{\nu}_k \tag{3.23a}$$

$$\bar{\nu}_k = y_k - H\hat{x}^s_{k/k-1} \tag{3.23b}$$

$$\hat{x}^s_{k/k} = \hat{x}^s_{k/k-1} + PH^T\Lambda^{-1}\bar{\nu}_k. \tag{3.23c}$$

Proof. The proof is similar to that in [7] for the one-step information delay case. Owing to the similarity of the calculations, some details will be omitted.

First we show that the control law

$$u_k = -L\hat{x}_{k/k-1} - M_k\nu_k \tag{3.24}$$

is admissible and optimizes J_{av}. Then we will show that the control law (3.22) is also admissible and yields the same optimal cost. Let $\tilde{x}_{k/k-1} = x_k - \hat{x}_{k/k-1}$. Under the control law (3.24), we obtain the closed-loop system

$$\hat{x}_{k+1/k} = (A - BL)\hat{x}_{k/k-1} + (K_k - BM_k)H\tilde{x}_{k/k-1} + (K_k - BM_k)e_k \tag{3.25}$$

$$\tilde{x}_{k+1/k} = (A - K_kH)\tilde{x}_{k/k-1} + (Cw_k - K_ke_k). \tag{3.26}$$

Define the matrices

$$A_k^c = \begin{bmatrix} A - BL & (K_k - BM_k)H \\ 0 & A - K_kH \end{bmatrix}$$

$$A^c = \begin{bmatrix} A - BL & (K - BM)H \\ 0 & A - KH \end{bmatrix}$$

$$N_k = \begin{bmatrix} 0 & K_k - BM_k \\ C & -K_k \end{bmatrix}$$

$$N = \begin{bmatrix} 0 & K - BM \\ C & -K \end{bmatrix}$$

$$W = \begin{bmatrix} I & G^T \\ G & R \end{bmatrix}.$$

W is the covariance of the process $\psi_k = [w_k^T \quad e_k^T]^T$. Using the fact that $\lim_{k\to\infty} \|A_k^c - A^c\| = 0$, $\lim_{k\to\infty} \|N_k - N\| = 0$, and A^c is asymptotically stable [7], we see that A_k^c is exponentially stable. Hence the process $\eta_k = [\hat{x}_{k/k-1}^T \quad \tilde{x}_{k/k-1}^T]^T$ has a correlation matrix $\Sigma_\eta(k) = E\eta_k\eta_k^T$ that satisfies

$$\Sigma_\eta(k+1) = A_k^c\Sigma_\eta(k)A_k^{c^T} + N_kWN_k^T. \tag{3.27}$$

By the time-varying Lyapunov lemma [1], $\lim_{k\to\infty}\Sigma_\eta(k) = \Sigma_\eta$ exists and is bounded. This implies that $E\|x_k\|^2$ is bounded for all k. Thus (3.24) is an admissible control law. Denote this control law by u'. Let

$$J_N(u) = E\left\{x_N^TSx_N + \sum_{k=0}^{N-1} \|Dx_k + Fu_k\|^2\right\}$$

where S is the unique positive semidefinite solution of CARE (3.20). Theorem 1 yields that $J_N(u) \geqq J_N(u')$. Furthermore, since for any admissible control law u,

$$\lim_{N\to\infty} \frac{1}{N} J_N(u) = J_{av}(u) \geqq J_{av}(u')$$

we see that u' is indeed optimal.

We now examine the closed-loop system under the time-invariant control law (3.22), denoted by u^*. Let $\beta_k = [\hat{x}_{k/k-1}^{s^T} \quad x_k^T - \hat{x}_{k/k-1}^{s^T}]^T$. Then

$$\beta_{k+1} = A^c\beta_k + N\psi_k. \tag{3.28}$$

Putting $\Sigma_\beta(k) = E\beta_k\beta_k^T$, we have immediately that $\lim_{k\to\infty}\Sigma_\beta(k) = \Sigma_\beta$ exists, and that

$$\Sigma_\beta = A^c\Sigma_\beta A^{c^T} + NWN^T. \tag{3.29}$$

Hence u^* is also admissible. Taking limit on both sides of (3.27), we see that $\Sigma_\eta = \Sigma_\beta$. A straightforward calculation similar to that in [7] now gives

$$J_{av}(u^*) = J_{av}(u').$$

Hence the control law u^*, given by (3.22), is also optimal, completing the proof.

The form of the optimal control is of course entirely expected given the results of Theorem 1 and standard LQG theory. In the next section, we shall use Theorem 2 to prove the explicit connection between LQG theory and minimum variance control for ARMAX systems with general delay.

4. Minimum variance control for SISO ARMAX systems. Consider the single-input single-output (SISO) ARMAX system

$$A(z^{-1})y_k = z^{-d}B(z^{-1})u_k + C(z^{-1})w_k \tag{4.1}$$

where

$$A(z^{-1}) = 1 + a_1z^{-1} + \cdots + a_nz^{-n}$$
$$B(z^{-1}) = b_d + b_{d+1}z^{-1} + \cdots + b_nz^{-n+d}$$
$$C(z^{-1}) = 1 + c_1z^{-1} + \cdots + c_nz^{-n}$$

and w_k is an independent and identically distributed zero-mean Gaussian sequence with $Ew_k^2 = \sigma^2$, z^{-1} is the backward shift, and $d \geqq 1$ is the system delay. Assume that the polynomials $B(q)$ and $C(q)$ have no roots in $|q| \leqq 1$. Then it is well known [14] that the optimal control law in the class of control laws of the form $u_k = \phi_k(y^k)$ which minimizes $\lim_{N\to\infty}(1/N)\sum_{k=0}^{N-1} Ey_k^2$ is given by

$$u_k = -\frac{G(z^{-1})}{B(z^{-1})F(z^{-1})} y_k \tag{4.2}$$

where the polynomials $F(z^{-1})$ and $G(z^{-1})$, of degrees $d-1$ and $n-1$, respectively, are determined by the equation

$$C(q)=A(q)F(q)+q^dG(q). \tag{4.3}$$

Furthermore, the closed-loop system is stable.

Let us reformulate the minimum variance control problem in the LQG framework. It is well known [6], [7] that the ARMAX description can be represented in innovation state space form as follows:

$$x_{k+1}=\begin{bmatrix} 0 & \cdots & 0 & -a_n \\ 1 & & \vdots & \\ 0 & & \vdots & \vdots \\ \vdots & & 0 & \\ 0 & \cdots & 1 & -a_1 \end{bmatrix} x_k+\begin{bmatrix} b_n \\ \vdots \\ b_d \\ 0 \\ \vdots \\ 0 \end{bmatrix} u_k+\begin{bmatrix} c_n-a_n \\ \vdots \\ c_1-a_1 \end{bmatrix} w_k = Ax_k+bu_k+cw_k \tag{4.4a}$$

$$y_k=[0 \quad \cdots \quad 0 \quad 1]x_k+w_k=hx_k+w_k. \tag{4.4b}$$

Observe that

$$hA^{j-1}b=0, \qquad j<d \tag{4.5a}$$

$$hA^{d-1}b=b_d. \tag{4.5b}$$

The cost criterion

$$J_{av}=\lim_{N\to\infty}\frac{1}{N}E\sum_{j=0}^{N-1}y_j^2=\lim_{N\to\infty}\frac{1}{N}E\sum_{j=0}^{N-1}(hx_j)^2+\sigma^2$$

so that J_{av} is minimized if and only if $\bar{J}_{av}=\lim_{N\to\infty}(1/N)\sum_{k=0}^{N-1}E(hx_k)^2$ is minimized.

The minimum variance control problem becomes that of LQG control with average cost per unit time criterion for the system described by (4.4a), (4.4b). Although the two approaches have been widely referred to as giving the same control law, an explicit proof cannot be found in the literature for the general delay case. In this section, we will prove this "folk theorem" using the general results of the previous section.

First, we observe that the minimum variance control problem is a singular control problem in LQG terminology since there is no control weighting. This singular problem can be converted to a nonsingular one as shown in [12], [18]. The control algebraic Riccati equation is given by

$$S=A^TSA-A^TSb(b^TSb)^{-1}b^TSA+h^Th \tag{4.6}$$

and it has been shown by Silverman [18] that the unique positive semidefinite solution in the minimum phase case is given by

$$S=\sum_{j=0}^{d-1}A^{j^T}h^ThA^j. \tag{4.7}$$

Furthermore the closed-loop system matrix $A-BL$ is stable. It is also well known [7], [15] that with $C(q)$ stable, the unique positive semidefinite solution to the filter algebraic Riccati equation

$$P=APA^T+\sigma^2cc^T-(APh^T+\sigma^2c)(hPh^T+\sigma^2)^{-1}(hPA^T+\sigma^2c^T)$$

is $P=0$. According to Theorem 2, the optimal control law is given by

$$u_k=-\Gamma^{-1}b^TS(A\hat{x}^s_{k/k-1}+K\bar{\nu}_k) \tag{4.8}$$

where $\hat{x}^s_{k/k-1}$ satisfies

(4.9a) $$\hat{x}^s_{k+1/k} = (A - ch)\hat{x}^s_{k/k-1} + bu_k + cy_k$$

(4.9b) $$\bar{\nu}_k = y_k - h\hat{x}^s_{k/k-1}.$$

Using (4.5) and (4.7) in (4.8) and (4.9), we can solve for the optimal controller transfer function from y to u to give

(4.10) $$u_k = -\frac{1}{b_d}\left\{hA^{d-1}(A-ch)\left[zI - \left(I - \frac{1}{b_d}bhA^{d-1}\right)(A-ch)\right]^{-1} \cdot \left[I - \frac{1}{b_d}bhA^{d-1}\right]c + hA^{d-1}c\right\}y_k.$$

To show that the state space LQG theory gives the same controller as the ARMAX minimum variance controller amounts to showing that (4.2) and (4.10) represents the same transfer function. Although (4.10) looks formidable, it in fact has a great deal of structure. This structure enables us to prove the following theorem.

THEOREM 3. *Assume that $d \geqq 1$ and that the polynomials $B(q)$ and $C(q)$ have no roots in $|q| \leqq 1$. Then the control law given in* (4.10) *minimizes J_{av} and is identical to the minimum variance control law* (4.2). *Furthermore, the closed-loop system is stable.*

Proof. Theorem 2 and the preceding calculations show that the control law (4.10) minimizes J_{av}, and that the closed-loop system is stable. It remains to show the equivalence of (4.2) and (4.10). Let

$$\bar{A} = A - ch$$
$$\bar{B} = \left(I - \frac{1}{b_d}bhA^{d-1}\right).$$

Then the controller of (4.10) is

(4.11) $$u_k = -\frac{1}{b_d}hA^{d-1}\{\bar{A}(zI - \bar{B}\bar{A})^{-1}\bar{B} + I\}cy_k.$$

By the matrix inversion lemma, the term inside the curly brackets can be simplified to give

(4.12) $$u_k = -\frac{1}{b_d}hA^{d-1}(I - z^{-1}\bar{A}\bar{B})^{-1}cy_k.$$

Let

(4.13) $$H(z^{-1}) = \frac{1}{b_d}hA^{d-1}(I - z^{-1}\bar{A}\bar{B})^{-1}c,$$

i.e., $H(z^{-1})$ is the controller transfer function. For the unit delay case $(d=1)$, the reader is referred to [5] or [16]. Here we concentrate on the case $d > 1$.

We make repeated use of the matrix inversion lemma in the following form: For any invertible matrix Z and vectors α, β, whenever the indicated inverse exists,

(4.14) $$(Z + \alpha\beta^T)^{-1} = Z^{-1} - \frac{Z^{-1}\alpha\beta^T Z^{-1}}{1 + \beta^T Z^{-1}\alpha}.$$

We also make use of the following result, proved in [15]:

(4.15) $$\frac{G(z^{-1})}{A(z^{-1})} = zh(zI - A)^{-1}A^{d-1}c.$$

We begin by noting that since $d > 1$, $hb = 0$. Hence

$$(4.16)\qquad (I - z^{-1}\bar{A}\bar{B})^{-1} = z\left[zI - A + ch + \frac{1}{b_d}AbhA^{d-1}\right]^{-1}.$$

Applying (4.14), with $zI - A + ch$ identified as Z, α as $(1/b_d)Ab$, and β^T as hA^{d-1}, we readily verify that

$$(4.17)\qquad \begin{aligned} b_dH(z^{-1}) &= hA^{d-1}z\left(zI - A + ch + \frac{1}{b_d}AbhA^{d-1}\right)^{-1}c \\ &= \frac{zhA^{d-1}(zI - A + ch)^{-1}c}{1 + hA^{d-1}(zI - A + ch)^{-1}(Ab/b_d)}. \end{aligned}$$

In the Appendix, the following two equations are proved:

$$(4.18)\qquad hA^{d-1}(zI - A + ch)^{-1}c = \frac{G(z^{-1})}{zC(z^{-1})}$$

$$(4.19)\qquad 1 + hA^{d-1}(zI - A + ch)^{-1}\frac{Ab}{b_d} = \frac{1}{b_d}\frac{B(z^{-1})}{A(z^{-1})}\frac{C(z^{-1}) - z^{-d}G(z^{-1})}{C(z^{-1})}.$$

On using (4.3), we see that (4.19) becomes

$$1 + hA^{d-1}(zI - A + ch)^{-1}\frac{Ab}{b_d} = \frac{1}{b_d}\frac{B(z^{-1})F(z^{-1})}{C(z^{-1})}.$$

Substituting (4.18) and (4.20) into (4.17), we find

$$b_dH(z^{-1}) = \frac{b_dG(z^{-1})}{B(z^{-1})F(z^{-1})}$$

which is the desired result.

Remark. The control law (4.10) is identical to (4.2) whether or not $B(q)$ is stable. However, if $B(q)$ is not stable, closed-loop stability no longer holds. The case where $B(q)$ is not stable but where the closed-loop system is required to be stable requires a somewhat different treatment using the stabilizing rather than optimizing solution of the Riccati equation. These results and extensions to the multivariable case will be reported elsewhere.

5. Conclusions. We have given an explicit solution in feedback form to the LQG problem with correlated dynamic and observation noise and no information delay. Certainty equivalence in its standard form is shown not to hold, but is shown to hold for an augmented system. The general results are applied to the minimum variance control problem for SISO ARMAX systems and the equivalence of the two approaches explicitly demonstrated. It is believed that these results fill an apparent gap in the LQG theory. Using these results, further connections between state space and ARMAX formulations for multivariable stochastic control problems, examined, for example, in [17], can be explored. Such investigations will be reported in forthcoming papers.

Appendix. The following equations will be proved:

$$(\text{A.1})\ (\text{i})\qquad hA^{d-1}(zI - A + ch)^{-1}c = \frac{G(z^{-1})}{zC(z^{-1})}$$

$$(\text{A.2})\ (\text{ii})\qquad 1 + hA^{d-1}(zI - A + ch)^{-1}\frac{Ab}{b_d} = \frac{1}{b_d}\frac{B(z^{-1})}{A(z^{-1})}\frac{[C(z^{-1}) - z^{-d}G(z^{-1})]}{C(z^{-1})}.$$

Proof of (i). Applying matrix inversion as in (4.14) to $(zI-A+ch)^{-1}$, we have

$$\begin{aligned} hA^{d-1}(zI-A+ch)^{-1}c &= h(zI-A)^{-1}A^{d-1}c - \frac{h(zI-A)^{-1}A^{d-1}ch(zI-A)^{-1}}{1+h(zI-A)^{-1}c}c \\ &= \frac{h(zI-A)^{-1}A^{d-1}c}{1+h(zI-A)^{-1}c}. \end{aligned}\tag{A.3}$$

It is easily verified that

$$1+h(zI-A)^{-1}c = \frac{C(z^{-1})}{A(z^{-1})}.\tag{A.4}$$

We finally obtain, on using (4.15), that

$$hA^{d-1}(zI-A+ch)^{-1}c = \frac{z^{-1}G(z^{-1})}{A(z^{-1})}\,\frac{A(z^{-1})}{C(z^{-1})} = z^{-1}\frac{G(z^{-1})}{C(z^{-1})}.$$

Proof of (ii). Applying matrix inversion as before, we have that

$$\begin{aligned} &hA^{d-1}(zI-A+ch)\frac{Ab}{b_d} \\ &\quad = h(zI-A)^{-1}A^{d-1}\frac{Ab}{b_d} - \frac{h(zI-A)^{-1}A^{d-1}ch(zI-A)^{-1}(Ab/b_d)}{1+h(zI-A)^{-1}c} \\ &\quad = \frac{[1 \quad z \quad \cdots \quad z^{n-1}]A(A^{d-1}b/b_d)}{z^nA(z^{-1})} \\ &\qquad - \frac{z^{-1}G(z^{-1})}{z^nA(z^{-1})C(z^{-1})}[1 \quad z \quad \cdots \quad z^{n-1}]\frac{Ab}{b_d} \end{aligned}\tag{A.5}$$

on using the structure of $h(zI-A)^{-1}$ and (4.15). Now observe that

$$\frac{1}{b_d}A^{d-1}b = \begin{bmatrix} 0 & \cdots & 0 & \frac{b_n}{b_d} & \frac{b_{n-1}}{b_d} & \cdots & \frac{b_{d+1}}{b_d} & 1 \end{bmatrix}^T\tag{A.6}$$

$$\frac{1}{b_d}Ab = \begin{bmatrix} 0 & \frac{b_n}{b_d} & \cdots & \frac{b_{d+1}}{b_d} & 1 & 0 & \cdots & 0 \end{bmatrix}^T\tag{A.7}$$

and

$$[1 \quad z \quad \cdots \quad z^{n-1}]A = [z \quad \cdots \quad z^{n-1} \quad -(a_1z^{n-1}+\cdots+a_n)].\tag{A.8}$$

Substituting (A.6)-(A.8) into (A.5), we obtain

$$\begin{aligned} hA^{d-1}(zI-A+ch)\frac{Ab}{b_d} &= \frac{-\sum_{i=1}^n a_iz^{n-1}+(1/b_d)\sum_{j=d+1}^n b_jz^{n+d-j}}{z^nA(z^{-1})} \\ &\quad - \frac{z^{-1}G(z^{-1})}{z^nA(z^{-1})C(z^{-1})}\,\frac{1}{b_d}\sum_{j=d}^n b_jz^{n-j+1}. \end{aligned}\tag{A.9}$$

Hence

$$\begin{aligned} 1+hA^{d-1}(zI-A+ch)\frac{Ab}{b_d} &= \frac{z^n+(1/b_d)\sum_{j=d+1}^n b_jz^{n+d-j}}{z^nA(z^{-1})} \\ &\quad - \frac{z^{-1}G(z^{-1})}{z^nA(z^{-1})C(z^{-1})}\,\frac{1}{b_d}\sum_{j=d}^n b_jz^{n-j+1} \\ &= \frac{1}{b_d}\,\frac{B(z^{-1})}{A(z^{-1})} - \frac{1}{b_d}z^{-d}\frac{G(z^{-1})B(z^{-1})}{C(z^{-1})A(z^{-1})} \\ &= \frac{1}{b_d}\,\frac{B(z^{-1})}{A(z^{-1})}\left[\frac{C(z^{-1})-z^{-d}G(z^{-1})}{C(z^{-1})}\right]. \end{aligned}$$

REFERENCES

[1] B. D. O. Anderson and J. B. Moore, *Detectability and stabilizability of time-varying discrete-time linear systems*, SIAM J. Control Optim., 19 (1981), pp. 20-32.

[2] K. J. Astrom, *Introduction to Stochastic Control Theory*, Academic Press, New York, 1970.

[3] K. J. Astrom and B. Wittenmark, *On self-tuning regulators*, Automatica, 9 (1973), pp. 185-199.

[4] D. P. Bertsekas, *Dynamic Programming and Stochastic Control*, Academic Press, New York, 1976.

[5] P. E. Caines, *Relationships between Box-Jenkins-Astrom control and Kalman linear regulator*, Proc. IEE, 119 (1972), pp. 615-620.

[6] P. E. Caines, *Linear Stochastic Systems*, John Wiley, New York, 1988.

[7] M. H. A. Davis and R. B. Vinter, *Stochastic Modelling and Control*, Chapman and Hall, New York, 1985.

[8] M. J. Grimble, *The design of stochastic optimal feedback control systems*, Proc. IEE, 125 (1978), pp. 1275-1284.

[9] ———, *Solution of the discrete-time stochastic optimal control problem in the z-domain*, Internat. J. Systems Sci., 10 (1979), pp. 1369-1390.

[10] ———, *Implicit and explicit* LQG *self-tuning controllers*, Automatica, 20 (1984), pp. 661-669.

[11] C. J. Harris and S. A. Billings, eds., *Self-tuning and Adaptive Control: Theory and Applications*, IEE publication, Peter Peregrinus, London, 1981.

[12] V. Kucera, *State space approach to discrete linear control*, Kybernetika, 8 (1972), pp. 233-251.

[13] ———, *Discrete Linear Control*, John Wiley, Chichester, 1979.

[14] P. R. Kumar and P. Varaiya, *Stochastic Systems: Estimation, Identification, and Adaptive Control*, Prentice-Hall, Englewood Cliffs, New Jersey, 1986.

[15] R. H. Kwong, *A simple characterization of optimal* ARMA *predictors*, Syst. Control Lett., 6 (1986), pp. 353-355.

[16] ———, *On the* LQG *problem with correlated noise and its relation to minimum variance control*, Systems Control Group Report No. 8620, University of Toronto, Toronto, Canada, Nov. 1986.

[17] U. Shaked and P. R. Kumar, *Minimum variance control of discrete time multivariable* ARMAX *systems*, SIAM J. Control Optim., 24 (1986), pp. 396-411.

[18] L. M. Silverman, *Discrete Riccati equations: Alternative algorithms, asymptotic properties, and system theory interpretations*, Control and Dynamic Systems, Vol. 12, C. T. Leondes, ed., Academic Press, New York, pp. 313-386.

[19] K. Warwick, *Optimal observers for* ARMA *models*, Internat. J. Control, 46 (1987), pp. 1493-1503.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 1, pp. 153-159, January 1991

008

NOTE ON THE CONVERGENCE OF SIMULATED ANNEALING ALGORITHMS*

U. FAIGLE† AND W. KERN†

Abstract. Generalizing the results of Faigle and Schrader [*Inform. Process. Lett.*, 27 (1988), pp. 189-194] a short inductive proof is given that shows that the stationary distributions of a simulated annealing algorithm converge to a distribution, where nonoptimal elements are generated with probability zero, provided that the "weak reversibility condition" of Hajek [*Math. Oper. Res.*, 13 (1988), pp. 311-329] holds.

Key words. simulated annealing, Markov process, combinatorial optimization, convergence

AMS(MOS) subject classifications. 90C40, 68J05

1. Introduction.

Simulated annealing was proposed by Metropolis et al. [Met53] as a Monte Carlo method for the evaluation of state equations in statistical mechanics. Its potential as a promising tool for approximately solving large-scale combinatorial optimization problems was later pointed out in [Kir83] and [Cer85].

Essentially, the method proceeds as follows. In order to generate an element of minimal weight in a set with some neighborhood structure, we iteratively choose a random neighbor of the current element. Whether this new element is accepted as the new current element depends on the outcome of a second probabilistic experiment, where the discriminating probability usually is based on the weights of the two elements under consideration and on a temperature parameter $t > 0$. Guided by thermodynamical analogies, we hope to generate an element of "lowest energy," i.e., of minimal weight, as $t \rightarrow 0$ (see, e.g., [Bur84], [Ros85], or [Gol86]).

We should make the scope of our investigation very clear. Any practical implementation of simulated annealing will, of course, try to include a feature that permits us to always keep track of the best solution discovered during the search procedure so far. From a theoretical point of view, it is obvious that such a feature guarantees that we encounter a globally best solution with probability eventually approaching one even in a pure random search (without any simulated annealing mechanism) (see also [Ani87]). Simulated annealing is a good deal more subtle in that it tries to steer the search procedure in the "right direction," i.e., it tries to eventually only generate solutions that are close to optimal with high probability.

However little is presently known about how to implement a simulated annealing algorithm so that the best results are achieved. Thus reports about practical success with simulated annealing are also skeptical (see, e.g., [Joh84]).

From a theoretical point of view, a simulated annealing algorithm can be modeled as an inhomogeneous Markov process, whose transition probabilities depend on the parameter t. In this framework, however, a rigorous analysis of the convergence behavior appears to be quite involved (see, e.g., [Haj88], [Chi88], [Tsi89]).

On the other hand, it is not too hard to show that for a large class of accepting probability functions, the stationary distributions of the homogeneous Markov processes associated with each fixed t converge with $t \rightarrow 0$ to a (generally unknown) limiting distribution, which, in fact, is also the limiting distribution of the inhomogeneous process if the temperature is lowered "slowly enough," where "slowly enough" roughly means so slowly that during the search procedure an optimal solution is encountered

* Received by the editors March 22, 1989; accepted for publication (in revised form) April 15, 1990.

† Faculty of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands.

at all with probability one (see Theorem 2 of [Ani87]). Somewhat surprisingly, this result is independent of the particular neighborhood structure.

With respect to the goal of simulated annealing, therefore, the question arises of whether the limiting distribution has the desirable property that nonoptimal elements only occur with probability zero. As it turns out, the affirmative answer to this question very much depends on the neighborhood structure. Roughly speaking, it seems that this structure should be in a sense "symmetric." Thus an affirmative answer is given in [Lun86] for numerically symmetric neighborhods, while in [Fai88] and [Con88] this condition is relaxed to combinatorially symmetric neighborhoods.

The purpose of this paper is to provide a short argument for an even more general model of symmetry, namely, the "weak reversibility" of Hajek [Haj88] (see also [Con89]), which stipulates only that an element i be reachable from an element j without exceeding weight c whenever j is reachable from i without exceeding weight c.

An incidental remark may be in order. Experimental results indicate that an appropriate neighborhood structure might be even more important for a practically successful simulated annealing algorithm than the choice of the temperature levels (see, e.g., [Fai88a], [Gol88]). To make this more precise, however, is beyond the current theoretical understanding of simulated annealing.

We discuss simulated annealing algorithms in terms of Markov processes in § 2. The proof of our main result is given in § 3.

2. Simulated annealing and Markov processes. In this section, we give a short description of simulated annealing algorithms in terms of Markov processes.

Let $S=\{1,2,\cdots,n\}$ be a finite set whose elements are weighted with real numbers c_i $(i\in S)$. Furthermore, let $A=(a_{ij})$ be a *stochastic* $(n\times n)$-matrix, i.e., $a_{ij}\geqq 0$ and

$$\sum_{j=1}^{n} a_{ij}=1 \qquad (i=1,\cdots,n).$$

With A we associate the (directed) *neighborhood graph* $G=G(A)$, whose vertices are the elements of S and whose edges are those pairs (i,j) of vertices satisfying

$$i=j \quad \text{or} \quad a_{ij}>0.$$

We will assume that G is strongly connected or, equivalently, that the homogeneous Markov process with transition matrix A is *irreducible.* In particular, there is a unique *stationary distribution* of A, i.e., a probability distribution $\pi=(\pi_1,\cdots,\pi_n)\in\mathbb{R}^n$ on S such that

$$\pi A=\pi.$$

Note that the rows of the matrix

$$\Pi=\lim_{k\to\infty} A^k$$

are identical to π (provided the limit exists).

Consider a family of *accepting probability functions*, i.e., functions $f_t:\mathbb{R}\times\mathbb{R}\to\mathbb{R}$ such that for every choice of $c_1', c_2',\cdots,c_n'\in\mathbb{R}$, there exists some t_0 with the properties

(i) $0<f_t(c_i',c_j')\leqq 1$ if $c_j'\geqq c_i'$ and $t\leqq t_0$

(ii) $f_t(c_i',c_j')f_t(c_j',c_k')=f_t(c_i',c_k')$ if $c_k'\geqq c_j'\geqq c_i'$ and $t\leqq t_0$,

where the index $t>0$ is a *temperature* parameter.

Examples are provided by the standard exponential accepting probabilities

$$f_t(c_i',c_j')=\exp\{-(c_j'-c_i')/t\},$$

or by the functions

$$f_t(c_i',c_j')=\frac{c_i'}{c_j'}\cdot t^{(c_j'-c_i')}.$$

Relative to A and f_t, we define the stochastic matrix $P(t)=(p_{ij}(t))$ via

$$p_{ij}(t)=\begin{cases} a_{ij} & \text{if } c_i \geqq c_j, \\ a_{ij}f_t(c_i, c_j) & \text{if } c_i < c_j, \\ 1-\sum_{m\neq i} p_{im}(t) & \text{if } i=j. \end{cases}$$

Note that the matrices A and $P(t)$ possess the same neighborhood graph G!

A *simulated annealing algorithm* relative to (A, f_t) is a random walk on G according to a chosen *annealing schedule* $t_1 \geqq t_2 \geqq \cdots \geqq t_l \geqq \cdots$ with

$$\lim_{l\to\infty} t_l = 0,$$

that is, a discrete inhomogeneous Markov process whose transition probabilities in the lth step are given by the matrix $P(t_l)$.

It follows from Theorem 2 of [Ani87] that

$$\pi^* = \lim_{t\to 0} \pi(t)$$

exists, where $\pi(t)$ is the stationary distribution of $P(t)$, for a large class of accepting probability functions, including the standard

$$f_t(x, y) = \exp\{-(y-x)/t\}.$$

Moreover, π^* is the limiting distribution for the simulated annealing algorithm if the temperature is lowered "slowly enough" (see also [Haj88]).

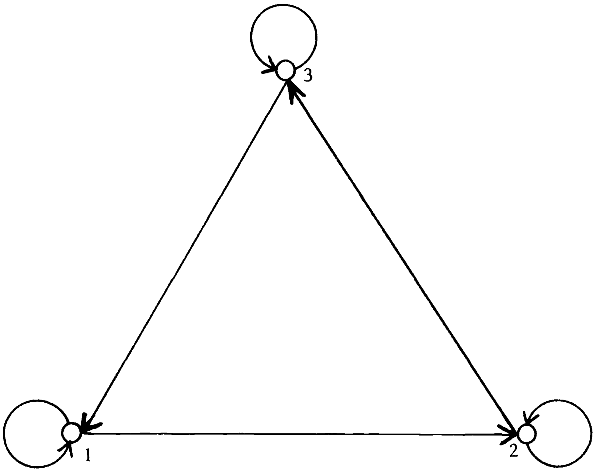
In this note, we are not interested in the choice of annealing schedules, but concentrate on the problem of whether

$$\lim_{t\to 0} \pi_s(t) = 0$$

if $s \in S$ is not *optimal*, i.e., if $s \notin S^0 = \{s \in S\colon c_r \geqq c_s \text{ for all } r \in S\}$.

Our next example shows that such a strong convergence property cannot be expected if we do not impose further conditions on the neighborhood structure A.

Example. Let $S=\{1, 2, 3\}$ and let the neighborhood matrix $A=(a_{ij})$ be defined by $a_{12}=a_{23}=a_{31}=1$. This yields $G(A)$ as follows:



Relative to the cost function $c_i = i$ $(i=1, 2, 3)$, we have $2 \notin S^0$. Yet, it is easy to see that the standard exponential accepting function yields the stationary probability:

$$\pi_2(t) = \frac{1}{2+e^{-1/t}}$$

and hence $\lim_{t\to 0} \pi_2(t) = \frac{1}{2} \neq 0$.

We will, therefore, assume that our neighborhood structure satisfies the "weak reversibility condition" relative to the weights c_i, which was introduced by Hajek in [Haj88]. To be more precise, we will assume that the neighborhood graph G satisfies the following condition relative to $c_1, \cdots, c_n$:

(WR) Each connected component of $G(c_i)$ is strongly connected.

Here $G(c_i)$ is the graph induced by G on the vertex set $S(c_i) = \{j \in S: c_j \leqq c_i\}$. Also recall that a directed graph is said to be *strongly connected* if every pair (u, v) of vertices is connected by a directed path from u to v.

We can now state our main result.

THEOREM 1. *If* (WR) *holds for* G, *then there exists a constant* K *and some* $t_0 > 0$ *such that*

$$\pi_r(t) \leqq K f_t(c_s, c_r) \pi_s(t)$$

whenever $t \leqq t_0$, $c_r > c_s$.

COROLLARY 2. *Assume that* (WR) *holds for* G *and that* $\lim_{t\to 0} f_t(c_s, c_r) = 0$ *for all* $c_r > c_s$. *Then*

$$\lim_{t\to 0} \pi_s(t) = 0$$

for all $s \in S \backslash S^0$.

To get an idea for the proof of Theorem 1, let us consider the case of the exponential accepting function

$$f_t(x, y) = \exp\{-(y-x)/t\}.$$

Suppose we knew that the associated stationary distributions $\pi(t)$ were, in a sense, "generalized Maxwell-Boltzmann," i.e., of the form

$$\pi_s(t) = Z(t) \cdot k_s \cdot e^{-c_s/t},$$

where $Z(t) = (\sum k_s e^{-c_s/t})^{-1}$ is a normalizing factor and the k_s's are constants depending only on s. Then, of course, Theorem 1 would follow immediately.

While we suspect that, under (WR), the stationary distributions are indeed "generalized Maxwell-Boltzmann," it turns out that Theorem 1 can be derived without first establishing a special form of $\pi(t)$. It suffices to show that the probabilities in question can be bounded from above by probabilities of the Maxwell-Boltzmann type. We will prove this in the next section.

3. Proof of the main result. For the proof of the main result, we replace the matrix A by a family $\{A(t)\}$ of stochastic matrices satisfying the ε-condition:

(EC) There exists an $\varepsilon > 0$ such that for each t and matrix $A(t) = (a_{ij}(t))$,

$$a_{ij}(t) \neq 0 \quad \text{implies } a_{ij}(t) \geqq \varepsilon$$

whenever $i \neq j$, and all $A(t)$'s have the same neighborhood graph.

For example, $A(t) := A$, for all t, gives rise to such a family.

With the notation as in the previous section, we can now formulate our main results as follows.

THEOREM 3. *Let* $\{A(t)\}$ *be a family of irreducible stochastic matrices satisfying the conditions* (EC) *and* (WR). *Then there exists a constant* K *and some* $t_0 > 0$ *such that*

$$\pi_r(t) \leqq K f_t(c_s, c_r) \pi_s(t)$$

whenever $t \leqq t_0$, $c_r > c_s$.

Proof. The proof is by induction on

$$m = |S| + |\{c_i : i \in S\}|,$$

i.e., the size of the ground set plus the number of distinct objective function values. For $m = 2$, there is nothing to prove. Thus assume $m \geqq 3$ and the theorem holds for values smaller than m. Now, either there exist two optimal solutions i and $j \in S^0$ with $a_{ij}(t) > 0$ (Case i below) or not (Case ii below). In the first case we modify the chain by "glueing" i and j into a single new optimal solution. This reduces the size of the ground set and hence our inductive assumption applies to the modified problem. In the second case, we "shift" all optimal solutions to the "second-to-best" level of the objective function value. (Note that such a "second-to-best" level must exist, since otherwise, in view of Case ii, the neighborhood graph would be disconnected.) This reduces the number of "level sets," i.e., the number of distinct weight function values, and hence, again, the inductive assumption applies to the modified problem. In both cases we show that the inductive assumption for the (smaller) modified problem implies that the claim also holds for the original problem. We now describe the above arguments in more detail.

Case i. There exist two distinct optimal elements $i, j \in S^0$ such that $a_{ij}(t) \geqq \varepsilon > 0$.

Because G is strongly connected, we have $\pi_i(t) > 0$ and $\pi_j(t) > 0$ for all $t > 0$. Thus

$$\lambda(t) = \frac{\pi_j(t)}{\pi_i(t) + \pi_j(t)}$$

is well defined.

This allows us to identify i and j as follows. For each matrix $A(t)$, we carry out the following construction:

(a) multiply the ith row by $\lambda(t)$
(b) multiply the jth row by $1 - \lambda(t)$
(c) add the new jth row to the new ith row
(d) delete the jth row
(e) add the new jth column to the new ith column
(f) delete the jth column.

It is clear that the resulting $(n-1) \times (n-1)$-matrices $\hat{A}(t)$ are stochastic and the underlying neighborhood graph $\hat{G}$ arises by contracting the edge (i, j) in G. Hence condition (WR) obviously holds for $\hat{A}(t)$.

Next let us verify that the family $\hat{A}(t)$ satisfies an ε-condition (EC). This amounts to showing that the multipliers $\lambda(t)$ and $1 - \lambda(t)$ can be bounded from below by some term which is independent of t. First note that

$$\pi_j(t) \geqq p_{ij}(t)\pi_i(t) = a_{ij}(t)\pi_i(t) \geqq \varepsilon\pi_i(t) \qquad (t > 0).$$

Thus

$$1 - \lambda(t) = \frac{\pi_j(t)}{\pi_i(t) + \pi_j(t)} \geqq \frac{\varepsilon}{1 + \varepsilon} \geqq \frac{\varepsilon}{2}.$$

Furthermore, since G satisfies (WR), we conclude that there exists a path from j to i in G, all of whose vertices are optimal. By multiplying the transition probabilities along this path, we get

$$\pi_i(t) \geqq \varepsilon^{n-1}\pi_j(t) \qquad (t > 0).$$

Thus

$$\lambda(t) = \frac{\pi_i(t)}{\pi_i(t) + \pi_j(t)} \geqq \frac{\varepsilon^{n-1}}{1 + \varepsilon^{n-1}} \geqq \frac{\varepsilon^{n-1}}{2}.$$

Hence both $\lambda(t)$ and $1-\lambda(t)$ can be bounded from below by $\varepsilon^{n-1}/2$. From this it is obvious that the family $\hat{A}(t)$ satisfies an ε-condition (EC) with $\hat{\varepsilon} := \varepsilon^n/2$.

Summarizing, we have shown that $\hat{A}(t)$ satisfies all conditions of Theorem 1. Hence, by induction, we conclude that the claim holds for $\hat{A}(t)$, i.e., there exist a constant $\hat{K}$ and some $\hat{t}_0 > 0$ such that

$$\hat{\pi}_r(t) \leqq \hat{K} f_t(c_s, c_r)\hat{\pi}_s(t)$$

holds for the stationary distribution $\hat{\pi}$ of the modified chain, whenever $t < t_0$ and $c_r > c_s$.

On the other hand, however, it is obvious that the stationary distribution $\hat{\pi}$ of the modified chain $\hat{P}(t)$ must be related to $\pi(t)$ via

$$\hat{\pi}_i(t) = \pi_i(t) + \pi_j(t),$$
$$\hat{\pi}_k(t) = \pi_k(t) \quad \text{for } k \neq i, j.$$

But this further implies that Theorem 1 also holds for the original chain with $\hat{K} = K \cdot 2 \cdot \varepsilon^{-(n-1)}$.

Case ii. For every two distinct $i, j \in S^0$, $a_{ij}(t) = 0$.

The idea for settling this case consists in reducing the number of different values assumed by the weights c_i. Thus we define

$$\delta := \min\{c_i : i \in S \backslash S^0\} - \min\{c_j : j \in S^0\}$$

and consider the weights $\hat{c}_i$ given by

$$\hat{c}_i = \begin{cases} c_i + \delta & \text{if } i \in S^0, \\ c_i & \text{if } i \in S\backslash S^0. \end{cases}$$

By induction, we know that the stationary distribution $\hat{\pi}(t)$ of the associated matrix $\hat{P}(t)$ satisfies

$$\hat{\pi}_r(t) \leqq \hat{K} f_t(\hat{c}_s, \hat{c}_r)\hat{\pi}_s(t) \quad \text{whenever } \hat{c}_r > \hat{c}_s,$$

for a suitable constant $\hat{K}$.

Now consider the vector $\bar{\pi}(t)$ with components

$$\bar{\pi}_i(t) = \begin{cases} \pi_i(t) & \text{if } i \in S\backslash S^0, \\ f_t(c_i, c_i + \delta)\pi_i(t) & \text{if } i \in S^0. \end{cases}$$

We claim that $\bar{\pi}(t)\hat{P}(t) = \bar{\pi}(t)$, i.e., up to normalizing, $\bar{\pi}(t)$ is the stationary distribution of $\hat{P}(t)$.

To verify the claim, first recall that

$$\sum_{i \in S} p_{ij}(t)\pi_i(t) = \pi_j(t).$$

Using the fact that

$$\hat{p}_{ij}(t) = 0 = p_{ij}(t) \quad \text{if } i, j \in S^0, i \neq j$$

and the multiplicative property

$$f_t(c_i, c_i + \delta) f_t(\hat{c}_i, \hat{c}_m) = f_t(c_i, c_m) \quad \text{if } i \in S^0, m \in S\backslash S^0,$$

we get the following:

For $i \in S^0$,

$$\begin{aligned} \hat{p}_{ii}(t)\bar{\pi}_i(t) &= \left[1 - \sum_{m \notin S^0} \hat{p}_{im}(t)\right] f_t(c_i, c_i + \delta)\pi_i(t) \\ &= \left[f_t(c_i, c_i + \delta) - \sum_{m \notin S^0} p_{im}(t)\right]\pi_i(t) \\ &= \left[\bar{\pi}_i(t) - \sum_{m \notin S^0} p_{im}(t)\right]\pi_i(t). \end{aligned}$$

For $i \notin S^0$, j arbitrary,

$$\hat{p}_{ij}(t)\bar{\pi}_j(t) = p_{ij}(t)\pi_j(t).$$

Thus

$$\sum_{i \in S} \hat{p}_{ij}(t)\bar{\pi}_i(t) = \bar{\pi}_j(t).$$

Writing $z(t) = \sum_{i \in S} \bar{\pi}_j(t)$, we therefore conclude that

$$\hat{\pi}(t) = z^{-1}(t)\bar{\pi}(t)$$

is the (unique) stationary distribution of $\hat{P}(t)$.

But this implies that the theorem holds for $P(t)$. Indeed, if $c_r > c_s$, then we have

$$\pi_r(t) \leqq \hat{K}f_t(c_r, c_s)\pi_s(t)$$

for $s \notin S^0$ and

$$\pi_r(t) \leqq \hat{K}f_t(\hat{c}_s, c_r)\bar{\pi}_s(t) = \hat{K}f_t(c_r, c_s)\pi_s(t)$$

for $s \in S^0$. □

REFERENCES

[Ani85] S. Anily and A. Federgruen, *Ergodicity in parametric non-stationary Markov chains: applications to simulated annealing methods*, working paper, Columbia University, New York, 1985.

[Ani87] ———, *Simulated annealing methods with general acceptance probabilities*, J. Appl. Probab., 24 (1987), pp. 657-667.

[Bur84] R. E. Burkard and F. Rendl, *A thermodynamically motivated simulation procedure of combinatorial optimization problems*, European J. Oper. Res., 17 (1984), pp. 169-174.

[Cer85] V. Cerny, *Thermodynamical approach to the traveling salesman problem: An efficient solution*, J. Optim. Theory Appl., 45 (1985), pp. 41-51.

[Chi88] T. Chiang and Y. Chow, *On eigenvalues and annealing rates*, Math. Oper. Res., 13 (1988), pp. 508-511.

[Con88] D. P. Connors and P. R. Kumar, *Balance of recurrence orders in time-inhomogeneous Markov chains with applications to simulated annealing*, Probab. Engrg. Inform. Sci., 2 (1988), pp. 157-184.

[Con89] ———, SIAM J. Comput., 27 (1989), pp. 440-461.

[Fai88] U. Faigle and R. Schrader, *On the convergence of stationary distributions in simulated annealing algorithms*, Inform. Process. Lett., 27 (1988), pp. 189-194.

[Fai88a] ———, *Simulated Annealing—Eine Fallstudie*, Angew. Inform., 6 (1988), pp. 259-264.

[Gol86] B. L. Golden and C. C. Skiscim, *Using simulated annealing to solve routing and location problems*, Nav. Res. Logist. Quart., 33 (1986), pp. 261-279.

[Gol88] L. Goldstein and M. Waterman, *The neighborhood size problem in simulated annealing*, Amer. J. Math. Management Sci., 8 (1988), pp. 409-423.

[Haj88] B. Hajek, *Cooling schedules for optimal annealing*, Math. Oper. Res., 13 (1988), pp. 311-329.

[Joh84] D. S. Johnson, C. R. Argon, L. A. McGeoch, and C. Schevon, *Optimization by simulated annealing: An experimental evaluation*, preprint, AT&T Bell Laboratories, Murray Hill, NJ, 1984.

[Kir83] S. Kirkpatrick, S. D. Gelatt jr., and M. P. Vecchi, *Optimization by simulated annealing*, Science, 220 (1983), pp. 671-680.

[Lun86] M. Lundy and A. Mees, *Convergence of the annealing algorithm*, Math. Programming, 34 (1986), pp. 111-124.

[Met53] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *Equations of state calculations by fast computing machines*, J. Chem. Phys., 21 (1953), pp. 1087-1092.

[Ros85] Y. Rossier, M. Troyon, and T. Liebling, *Probabilistic exchange algorithms and euclidean traveling salesman problems*, preprint, Dept. de Mathématiques, EPF, Lausanne, 1985.

[Tsi89] J. N. Tsitsiklis, *Markov Chains with rare transitions and simulated annealing*, Math. Oper. Res., 14 (1989), pp. 70-90.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 1, pp. 160-184, January 1991

THE SINGULAR H_∞ CONTROL PROBLEM WITH DYNAMIC MEASUREMENT FEEDBACK*

A. A. STOORVOGEL†

Abstract. This paper is concerned with the H_∞ problem with measurement feedback. The problem is to find a dynamic feedback from the measured output to the control input such that the closed-loop system has an H_∞ norm strictly less than some a priori given bound γ and such that the closed-loop system is internally stable. Necessary and sufficient conditions are given under which such a feedback exists. The only assumption that must be made is that there are no invariant zeros on the imaginary axis for two subsystems. Contrary to recent publications no assumptions are made on the direct feedthrough matrices of the plant. It turns out that this problem can be reduced to an almost disturbance decoupling problem with measurement feedback and internal stability, i.e., the problem in which we can make the H_∞ norm *arbitrarily small.*

Key words. quadratic matrix inequality, Riccati equation, almost disturbance decoupling, measurement feedback, internal stability

AMS(MOS) subject classifications. 93B27, 93B50, 93C05, 93C35, 93C45, 93C60

1. Introduction.

After the original formulation of the H_∞ problem in [22] much work has been done on the solution of this problem. Initially almost all the work was done in a mixture of time-domain and frequency-domain techniques (see [1], [4], [5]). In the last few years two new methods have evolved: the polynomial approach (see [9]) and a time-domain approach (see [2], [8], [12], [13], [20]).

This paper handles the problem in the time domain. This has the advantage that we directly obtain an upper bound on the necessary dynamic order of the controller, namely, the dynamic order of the original plant. A similar result was obtained in [10] and [11] using frequency domain techniques. Moreover, in our opinion, the results are more intuitive.

In the above-mentioned literature it was assumed that there are no invariant zeros on the imaginary axis and that the direct feedthrough matrices of the plant are nonsingular. In literature two methods have been proposed to tackle the H_∞ problem without these assumptions:

- Apply a small perturbation on the output matrices such that these assumptions are satisfied for the perturbed system. Then solve the H_∞ problem for the perturbed system. If the perturbation satisfies some prerequisites then a controller works for the original system if it works for the perturbed system. However, we do not know a priori how large the perturbations are allowed to be. Hence if for a certain perturbation no suitable controller exists, then we are not sure whether or not a suitable controller exists for a smaller perturbation (see [19]).
- Apply a transformation in the frequency domain:
$$G(s) \to \tilde{G}(s) := G\left(\frac{s+\varepsilon}{1+\varepsilon s}\right) \qquad (\varepsilon > 0).$$
If we can find a suitable controller for the original system, then we can find a controller for the transformed plant for ε small enough. Vice versa, if for some ε there exists a suitable controller for the transformed system then the same

* Received by the editors June 13, 1989; accepted for publication (in revised form) January 25, 1990.

† Department of Mathematics and Computing Science, Eindhoven University of Technology, P.O. Box 513, 5600 MB Eindhoven, The Netherlands.

controller is a suitable controller for the original system. This approach has the same disadvantage as the previous one since it is not clear how small we should choose ε. Another problem is that we still must make the assumption that the transfer matrix from control input to output is left invertible as a rational matrix (see [15]).

Recently, in the case of state feedback, a method of handling the singularity of the direct feedthrough matrix (see [18]) without the above-mentioned disadvantages was proposed. In the present paper we shall develop a method of handling these singularities in the case of measurement feedback. Our results reduce to the known results in [2] and [20] in case these singularities do not occur.

The necessary and sufficient conditions under which there exists an internally stabilizing dynamic compensator which makes the H_∞ norm strictly less than some a priori given bound γ are formulated in a way that differs from those found in recent publications [2], [20]. In these papers the results are formulated in terms of two Riccati equations. However in the case where there are singularities of the direct feedthrough matrices, these Riccati equations do not exist. We have two quadratic matrix inequalities that replace the role of these Riccati equations. The solution of each of these quadratic matrix inequalities must satisfy rank conditions. Moreover, we have a condition which couples these two matrix inequalities. The spectral radius of the product of the two solutions of these matrix inequalities should be smaller than a certain a priori given upper bound. In the regular case the first rank condition together with the quadratic matrix inequality reduces to a Riccati equation and the second rank condition guarantees that it is a stabilizing solution of the Riccati equation.

The proof of our main result only uses the result for the state feedback H_∞ control problem. Our proof will use ideas used in [2] to solve the regular H_∞ problem with measurement feedback but is independent of the results in [2] and is entirely self-contained.

The outline of the paper is as follows. In § 2 we formulate the problem and present the main result. Moreover, we show that in the regular case and the state feedback case this result reduces to the known results in [2] and [18], respectively. In § 3 it is shown that the conditions for the existence of a suitable compensator as given in our main theorem are necessary. It is also shown that the problem of finding such a compensator is equivalent to finding such a compensator for another system, i.e., it is shown that any compensator which internally stabilizes this new system and makes the closed-loop H_∞ norm less than γ has the same properties when applied to the original system and vice versa. This new system has some desirable properties and using these properties in § 4, it is shown that for this new system we can even make the H_∞ norm arbitrarily small. In § 5 a method for finding the desired compensator is discussed. We finish in § 6 with some concluding remarks. The proofs of § 3 are given in Appendix B since they are rather technical. Appendix A introduces a number of suitably chosen bases and some of the properties the system matrices have in these new bases. These will be needed in Appendix B.

2. Problem formulation and main results. We consider the linear, time-invariant, finite-dimensional system:

$$\Sigma: \begin{cases} \dot{x} = Ax + Bu + Ew, \\ y = C_1x + D_1w, \\ z = C_2x + D_2u, \end{cases} \tag{2.1}$$

where $x \in \mathcal{R}^n$ is the state, $u \in \mathcal{R}^m$ the control input, $w \in \mathcal{R}^l$ the unknown disturbance,

$y \in \mathcal{R}^p$ the measured output, and $z \in \mathcal{R}^q$ the unknown output to be controlled. A, B, E, C_1, C_2, D_1, and D_2 are matrices of appropriate dimensions. We would like to minimize the effect of the disturbance w on the output z, using the measured output y, by finding an appropriate control input u. More precisely, we seek a *dynamic* compensator F described by the following equations:

$$\Sigma_F : \begin{cases} \dot{p} = Kp + Ly, \\ u = Mp + Ny, \end{cases} \tag{2.2}$$

such that after applying the feedback $u = Fy$ in the system (2.1), the resulting closed-loop system, whose transfer matrix is denoted by G_F, is internally stable and has minimal H_∞ norm, i.e., such that

$$\|G_F\|_\infty := \sup_{\omega \in \mathcal{R}} \sigma[G_F(i\omega)] \tag{2.3}$$

is minimized over all possible dynamic feedback laws F that make the closed-loop system internally stable. Here $\sigma[M]$ denotes the largest singular value of the matrix M. Internally stable means that when $w \equiv 0$ then for every initial state of the system and controller the state of the system and controller in the interconnection both converge to zero as $t \to \infty$. If the controller is given by (2.2) and the system is given by (2.1) this is equivalent to requiring that the following matrix is asymptotically stable:

$$\begin{pmatrix} A + BNC_1 & BM \\ LC_1 & K \end{pmatrix}. \tag{2.4}$$

Although this is our ultimate goal, in this paper we shall derive necessary and sufficient conditions under which we can find a dynamic feedback law which makes the resulting H_∞ norm of the closed-loop system strictly less than some a priori given bound γ and such that the resulting closed-loop system is internally stable.

A central role in our study of the problem above will be played by the *quadratic matrix inequality*. For any $\gamma > 0$ and matrix $P \in \mathcal{R}^{n \times n}$ we define the following matrix:

$$F_\gamma(P) := \begin{pmatrix} A^TP + PA + C_2^TC_2 + \gamma^{-2}PEE^TP & PB + C_2^TD_2 \\ B^TP + D_2^TC_2 & D_2^TD_2 \end{pmatrix}. \tag{2.5}$$

If $F_\gamma(P) \geqq 0$, we say P is a solution of the quadratic matrix inequality at γ. We also define a dual version of this quadratic matrix inequality. For any $\gamma > 0$ and matrix $Q \in \mathcal{R}^{n \times n}$ we define the following matrix:

$$G_\gamma(Q) := \begin{pmatrix} AQ + QA^T + EE^T + \gamma^{-2}QC_2^TC_2Q & QC_1^T + ED_1^T \\ C_1Q + D_1E^T & D_1D_1^T \end{pmatrix}. \tag{2.6}$$

If $G_\gamma(Q) \geqq 0$, we say that Q is a solution of the dual quadratic matrix inequality at γ. In addition to these two matrices we define two polynomial matrices, whose role is again completely dual:

$$L_\gamma(P, s) := [sI - A - \gamma^2EE^TP \quad -B], \tag{2.7}$$

$$M_\gamma(Q, s) := \begin{bmatrix} sI - A - \gamma^{-2}QC_2^TC_2 \\ -C_1 \end{bmatrix}. \tag{2.8}$$

We note that $L_\gamma(P, s)$ is the controllability pencil associated with the system:

$$\dot{x} = (A + \gamma^{-2}EE^TP)x + Bu, \tag{2.9}$$

while $M_\gamma(Q, s)$ is the observability pencil associated with the system:

$$\dot{x} = (A + \gamma^{-2} Q C_2^T C_2)x, \qquad y = -C_1 x. \tag{2.10}$$

We define the following two transfer matrices which again play a dual role:

$$G(s) := C_2(sI - A)^{-1}B + D_2, \tag{2.11}$$

$$H(s) := C_1(sI - A)^{-1}E + D_1. \tag{2.12}$$

In the formulation of our main result we also require the concept of *invariant zero* of the system $\Sigma = (A, B, C, D)$. These are all $s \in \mathscr{C}$ such that

$$\operatorname{rank} \begin{pmatrix} sI - A & -B \\ C & D \end{pmatrix} < \operatorname{normrank} \begin{pmatrix} sI - A & -B \\ C & D \end{pmatrix}. \tag{2.13}$$

Here "normrank" denotes the rank of a matrix as a matrix with entries in the field of rational functions. Moreover let $\mathscr{C}^+ (\mathscr{C}^0, \mathscr{C}^-)$ denote all $s \in \mathscr{C}$ such that Re $s > 0$ (Re $s = 0$, Re $s < 0$). Finally, let $\rho(M)$ denote the spectral radius of the matrix M. We are now in the position to formulate our main result.

THEOREM 2.1. *Consider the system* (2.1). *Assume that the systems* (A, B, C_2, D_2) *and* (A, E, C_1, D_1) *have no invariant zeros in* $\mathscr{C}^0$. *Then the following two statements are equivalent*:

(i) *There exists a linear, time-invariant, finite-dimensional dynamic compensator* F *of the form* (2.2) *such that by applying* $u = Fy$ *in* (2.1) *the resulting closed-loop system, with transfer matrix* G_F, *is internally stable and has* H_∞ *norm less than* γ, *i.e.,* $\|G_F\|_\infty < \gamma$.

(ii) *There exist positive semidefinite solutions* P, Q *of the quadratic matrix inequalities* $F_\gamma(P) \geqq 0$ *and* $G_\gamma(Q) \geqq 0$ *satisfying* $\rho(PQ) < \gamma^2$, *such that the following rank conditions are satisfied*:

(1) rank $F_\gamma(P) = $ normrank G,

(2) rank $G_\gamma(Q) = $ normrank H,

(3) $\operatorname{rank} \begin{pmatrix} L_\gamma(P, s) \\ F_\gamma(P) \end{pmatrix} = n + \operatorname{normrank} G \quad \forall s \in \mathscr{C}^0 \cup \mathscr{C}^+$,

(4) $\operatorname{rank} (M_\gamma(Q, s) \; G_\gamma(Q)) = n + \operatorname{normrank} H \, \forall s \in \mathscr{C}^0 \cup \mathscr{C}^+$.

Remarks.

(i) Note that since $P \geqq 0$ and $Q \geqq 0$ the matrix PQ has only real and nonnegative eigenvalues.

(ii) The construction of a dynamic compensator satisfying (i) can be done according to the method as described in § 5. It turns out that it is always possible to find a compensator of the same dynamic order as the original plant.

(iii) By Corollary A5 we know that a solution P of the quadratic matrix inequality $F_\gamma(P) \geqq 0$ satisfying (1) and (3) is unique. By dualizing Corollary A5 it can also be shown that a solution Q of the dual quadratic matrix inequality $G_\gamma(Q) \geqq 0$ satisfying (2) and (4) is unique. The existence of P and Q can be checked via a state-space transformation and investigating a reduced order Riccati equation.

(iv) We shall prove this theorem only for the case $\gamma = 1$. The general result can then be easily obtained by scaling.

Before we prove this result we look more closely at the result for two special cases.

State feedback: $C_1 = I$, $D_1 = 0$. In this case we have $y = x$, i.e., we know the state of the system. The first matrix inequality $F_\gamma(P) \geqq 0$ together with rank conditions (1) and (3) does not depend on C_1 or D_1 so we cannot expect a simplification there. However $G_\gamma(Q)$ does get a special form:

$$G_\gamma(Q)=\begin{pmatrix} AQ+QA^T+EE^T+\gamma^{-2}QC_2^TC_2Q & Q \\ Q & 0 \end{pmatrix}. \tag{2.14}$$

Using this special form it can be easily seen that $G_\gamma(Q)\geqq 0$ if and only if $Q=0$. For the rank conditions it is interesting to investigate the normrank of H. We have

$$\text{normrank } H=\text{normrank }(sI-A)^{-1}E=\text{rank } E. \tag{2.15}$$

It can be easily checked, by using (2.15), that $Q=0$ satisfies rank conditions (2) and (4). The condition $\rho(PQ)<\gamma^2$ is trivially satisfied when $Q=0$. We find that in this case condition (ii) of Theorem 2.1 becomes:

There exists a positive semidefinite solution P of the matrix inequality $F_\gamma(P)\geqq 0$ such that the following two rank conditions are satisfied:

(1) rank $F_\gamma(P)=$ normrank G,

(2) rank $\begin{pmatrix} L_\gamma(P,s) \\ F_\gamma(P) \end{pmatrix}=n+\text{normrank } G \quad \forall s\in \mathscr{C}^0\cup\mathscr{C}^+$,

which is exactly the result obtained in [18].

Regular case: *D_1 surjective and D_2 injective.* In this case it can be shown, in the same way as in [18], that $F_\gamma(P)\geqq 0$ together with rank condition (1) is equivalent to the condition

$$A^TP+PA+C_2^TC_2+\gamma^{-2}PEE^TP-(PB+C_2^TD_2)(D_2^TD_2)^{-1}(B^TP+D_2^TC_2)=0.$$

The dual version of this proof can be applied to the dual matrix inequality $G_\gamma(Q)\geqq 0$ together with rank condition (2). These conditions turn out to be equivalent to the condition:

$$AQ+QA^T+EE^T+\gamma^{-2}QC_2^TC_2Q-(QC_1^T+ED_1^T)(D_1D_1^T)^{-1}(C_1Q+D_1E^T)=0.$$

The two remaining rank conditions (3) and (4) turn out to be equivalent with the requirement that the following two matrices are asymptotically stable:

$$A+\gamma^{-2}EE^TP-B(D_2^TD_2)^{-1}(B^TP+D_2^TC_2),$$
$$A+\gamma^{-2}QC_2^TC_2-(QC_1^T+ED_1^T)(D_1D_1^T)^{-1}C_1.$$

Together with the remaining condition $\rho(PQ)<\gamma^2$, we thus re-obtain exactly the conditions derived in [2] and [6].

3. Reduction of the original problem to an almost disturbance decoupling problem. In this section the implication (i)⇒(ii) in Theorem 2.1 will be proven. Moreover, in case the conditions (ii) of Theorem 2.1 are satisfied, we shall show that the problem of finding a suitable compensator F for the system (2.1) is equivalent to finding a suitable compensator F for a *new system* which has some very nice structural properties. In the next section the H_∞ problem for this new system will be tackled. *In the remainder of this paper we assume* $\gamma=1$. The general result can be easily obtained by scaling. Define $F(P)$, $G(Q)$, $L(P,s)$, and $M(Q,s)$ to be equal to $F_1(P)$, $G_1(Q)$, $L_1(P,s)$, and $M_1(Q,s)$, respectively.

LEMMA 3.1. *Assume that (A, B, C_2, D_2) and (A, E, C_1, D_1) have no invariant zeros on $\mathscr{C}^0$. If there exists a linear, time-invariant, finite-dimensional dynamic compensator F such that the resulting closed-loop system is internally stable and has H_∞ norm less than one, then the following two conditions are satisfied*:

(i) *There exists a solution* $P \geqq 0$ *of the quadratic matrix inequality* $F(P) \geqq 0$ *satisfying the following two rank conditions*:

(1) rank $F(P) =$ normrank G,

(2) $\operatorname{rank}\begin{pmatrix} L(P, s) \\ F(P) \end{pmatrix} = n + \operatorname{normrank} G \quad \forall s \in \mathscr{C}^0 \cup \mathscr{C}^+.$

(ii) *There exists a solution* $Q \geqq 0$ *of the dual quadratic matrix inequality* $G(Q) \geqq 0$ *satisfying the following two rank conditions*:

(1) rank $G(Q) =$ normrank H,

(2) $\operatorname{rank}(M(Q, s)\ G(Q)) = n + \operatorname{normrank} H \quad \forall s \in \mathscr{C}^0 \cup \mathscr{C}^+.$

Proof. Since there exists an internally stabilizing feedback which makes the H_∞ norm less than one for the problem with measurement feedback there certainly also exists an internally stabilizing feedback which makes the H_∞ norm less than one in the full information case, i.e., the case where both x and w are known. This implies, according to [18], that there exists a matrix P satisfying the conditions in (i). By dualization it can be easily shown that there also exists a matrix Q satisfying the conditions in (ii). □

Assume there exist P and Q satisfying the conditions in parts (i) and (ii) of Lemma 3.1. We make the following factorization of $F(P)$:

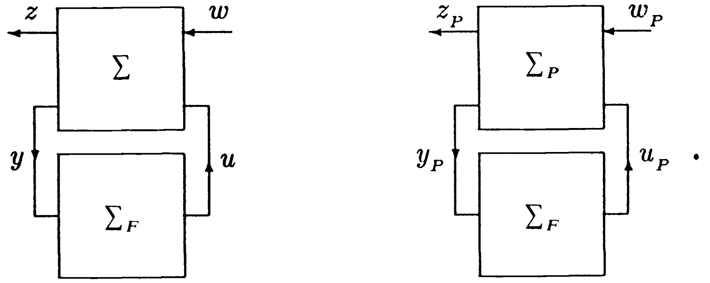
$$F(P) = \begin{pmatrix} C_{2,P}^T \\ D_P^T \end{pmatrix} (C_{2,P} \quad D_P) \tag{3.1}$$

where $C_{2,P}$ and D_P are matrices of suitable dimensions. This can be done since $F(P) \geqq 0$. We define the following system:

$$\Sigma_P : \begin{cases} \dot{x}_P = (A + EE^T P)x_P + Bu_P + Ew_P, \\ y_P = (C_1 + D_1 E^T P)x_P + D_1 w_P, \\ z_P = C_{2,P} x_P + D_P u_P. \end{cases} \tag{3.2}$$

LEMMA 3.2. *Let P satisfy Lemma* 3.1(i). *Moreover let an arbitrary linear time-invariant finite-dimensional compensator F be given, described by* (2.2). *Consider the following two systems, where the system on the left is the interconnection of* (2.1) *and* (2.2) *and the system on the right is the interconnection of* (3.2) *and* (2.2):

(3.3)



Then the following statements are equivalent:

(i) *The system on the left is internally stable and its transfer matrix from w to z has H_∞ norm less than one.*

(ii) *The system on the right is internally stable and its transfer matrix from w_P to z_P has H_∞ norm less than one.*

Proof. See appendix B for the proof. □

If for the original system (2.1) there exists an internally stabilizing, linear, time-invariant, finite-dimensional compensator such that the resulting closed-loop matrix has H_∞ norm less than one then, by applying Lemma 3.2, we know that the *same* compensator is internally stabilizing for the new system (3.2) and yields a closed-loop transfer matrix with H_∞ norm less than one. Hence if we consider for this new system the two quadratic matrix inequalities we know from Lemma 3.1 that there exist positive semidefinite solutions to these inequalities satisfying a number of rank conditions. We shall now formalize this in the following lemma. Define $A_P := (A + EE^TP)$ and $C_{1,P} := (C_1 + D_1E^TP)$. Then for arbitrary X and Y in $\mathscr{R}^{n\times n}$ we define the following matrices:

$$\bar{F}(X) := \begin{pmatrix} A_P^TX + XA_P + C_{2,P}^TC_{2,P} + XEE^TX & XB + C_{2,P}^TD_P \\ B^TX + D_P^TC_{2,P} & D_P^TD_P \end{pmatrix}, \tag{3.4}$$

$$\bar{G}(Y) := \begin{pmatrix} A_PY + YA_P^T + EE^T + YC_{2,P}^TC_{2,P}Y & YC_{1,P}^T + ED_1^T \\ C_{1,P}Y + D_1E^T & D_1D_1^T \end{pmatrix}, \tag{3.5}$$

$$\bar{L}(X, s) := [sI - A_P - EE^TX \quad -B], \tag{3.6}$$

$$\bar{M}(Y, s) := \begin{bmatrix} sI - A_P - YC_{2,P}^TC_{2,P} \\ -C_{1,P} \end{bmatrix}. \tag{3.7}$$

Moreover, we define two new transfer matrices:

$$\bar{G}(s) := C_{2,P}(sI - A_P)^{-1}B + D_P, \tag{3.8}$$

$$\bar{H}(s) := C_{1,P}(sI - A_P)^{-1}E + D_1. \tag{3.9}$$

LEMMA 3.3. *Let P and Q satisfy part* (i) *and part* (ii) *in Lemma* 3.1, *respectively. Assume* (A, B, C_2, D_2) *and* (A, E, C_1, D_1) *have no invariant zeros on* $\mathscr{C}^0$. *Then we have the following two results*:

(i) $X := 0$ *is a solution of the quadratic matrix inequality* $\bar{F}(X) \geqq 0$ *and satisfies the following two rank conditions*:

(1) rank $\bar{F}(X) =$ normrank $\bar{G}$,

(2) rank $\begin{pmatrix} \bar{L}(X, s) \\ \bar{F}(X) \end{pmatrix} = n +$ normrank $\bar{G} \quad \forall s \in \mathscr{C}^0 \cup \mathscr{C}^+$.

(ii) *There exist a matrix Y satisfying the quadratic matrix inequality* $\bar{G}(Y) \geqq 0$ *together with the following two rank conditions*:

(1) rank $\bar{G}(Y) =$ normrank $\bar{H}$,

(2) rank $(\bar{M}(Y, s) \quad \bar{G}(Y)) = n +$ normrank $\bar{H} \quad \forall s \in \mathscr{C}^0 \cup \mathscr{C}^+$,

if and only if $I - QP$ is invertible. Moreover, in that case $Y := (I - QP)^{-1}Q$ is the unique solution. This matrix Y is positive semidefinite if and only if

$$\rho(PQ) < 1. \tag{3.10}$$

Proof. See Appendix B for the proof. □

Proof of (i)⇒(ii) *in Theorem* 2.1. The first part can be obtained directly from Lemma 3.1. By Lemma 3.2 we know that also for the transformed system Σ_P there exists a dynamic compensator which internally stabilizes the system and makes the H_∞ norm less than one. By applying Lemma 3.1 to this new system, this implies that there exists a matrix $Y \geqq 0$ satisfying Lemma 3.3(ii). Hence by Lemma 3.3 we have (3.10) and therefore all the conditions in Theorem 2.1(ii) are satisfied. □

In the remainder of this section we assume that the conditions of Theorem 2.1(ii) are satisfied.

In order to prove the implication (ii)⇒(i) in Theorem 2.1 we transform the system (3.2) once again. This time, however, we use the dualized version of the original transformation. By Lemma 3.3 we know $Y = (I - QP)^{-1}Q \geqq 0$ satisfies $\bar{G}(Y) \geqq 0$. We factorize $\bar{G}(Y)$:

$$\bar{G}(Y) = \begin{pmatrix} E_{P,Q} \\ D_{P,Q} \end{pmatrix} (E_{P,Q}^T \quad D_{P,Q}^T) \tag{3.11}$$

where $E_{P,Q}$ and $D_{P,Q}$ are matrices of suitable dimensions. We define the following system:

$$\text{(3.12)}\qquad \Sigma_{P,Q}:\begin{cases} \dot{x}_{P,Q}=A_{P,Q}x_{P,Q}+B_{P,Q}u_{P,Q}+E_{P,Q}w,\\ y_{P,Q}=C_{1,P}x_{P,Q}\qquad\qquad\quad +D_{P,Q}w,\\ z_{P,Q}=C_{2,P}x_{P,Q}+D_Pu_{P,Q},\end{cases}$$

where

$$\text{(3.13)}\qquad A_{P,Q}:=A_P+YC_{2,P}^TC_{2,P},$$

$$\text{(3.14)}\qquad B_{P,Q}:=B+YC_{2,P}^TD_P.$$

By applying Lemma 3.3 to the system $\Sigma_{P,Q}$ with the corresponding matrix inequalities we note that $X_{P,Q}:=0$ and $Y_{P,Q}:=0$ satisfy the matrix inequalities and the corresponding rank conditions for this new system. It can be shown that this implies that

$$\text{(3.15)}\qquad \operatorname{rank}\begin{pmatrix} sI-A_{P,Q} & -B_{P,Q}\\ C_{2,P} & D_P\end{pmatrix}=n+\operatorname{rank}\,(C_{2,P}\quad D_P)\quad \forall s\in\mathscr{C}^0\cup\mathscr{C}^+$$

and

$$\text{(3.16)}\qquad \operatorname{rank}\begin{pmatrix} sI-A_{P,Q} & -E_{P,Q}\\ C_{1,P} & D_{P,Q}\end{pmatrix}=n+\operatorname{rank}\begin{pmatrix} E_{P,Q}\\ D_{P,Q}\end{pmatrix}\quad \forall s\in\mathscr{C}^0\cup\mathscr{C}^+.$$

By applying Lemma 3.2 and its dualized version the following corollary can be derived.

COROLLARY 3.4. *Let an arbitrary compensator F of the form* (2.2) *be given. The following two statements are equivalent*:

(i) *The compensator F when applied to the system Σ, described by* (2.1), *is internally stabilizing and the resulting closed-loop transfer matrix has H_∞ norm less than one.*

(ii) *The compensator F when applied to the system $\Sigma_{P,Q}$, described by* (3.12), *is internally stabilizing and the resulting closed-loop transfer matrix has H_∞ norm less than one.*

In the next section we shall show how to solve the H_∞ problem for a system satisfying the extra conditions (3.15) and (3.16). It turns out that for this new system we can even make the H_∞ norm arbitrarily small.

4. The solution of the almost disturbance decoupling problem. Assume that the following system is given:

$$\text{(4.1)}\qquad \Sigma:\begin{cases} \dot{x}=Ax+Bu+Ew\\ y=C_1x+\qquad\; D_1w,\\ z=C_2x+D_2u,\end{cases}$$

such that the following two conditions are satisfied:

$$\text{(4.2)}\qquad \begin{pmatrix} sI-A & -B\\ C_2 & D_2\end{pmatrix}=n+\operatorname{rank}\,(C_2\quad D_2)\quad \forall s\in\mathscr{C}^0\cup\mathscr{C}^+$$

and

$$\text{(4.3)}\qquad \begin{pmatrix} sI-A & -E\\ C_1 & D_1\end{pmatrix}=n+\operatorname{rank}\begin{pmatrix} E\\ D_1\end{pmatrix}\quad \forall s\in\mathscr{C}^0\cup\mathscr{C}^+.$$

From the previous section we know that if the conditions in part (ii) of Theorem 2.1 are satisfied then it is always possible to transform our system into a new system that satisfies the conditions (4.2) and (4.3). Moreover, if a compensator F given by (2.2) internally stabilizes this new system and makes the H_∞ norm of the resulting closed-loop transfer matrix smaller than one, then it does the same with the closed-loop system associated with the original system. In fact, we shall prove a stronger result.

THEOREM 4.1. *Assume system* (4.1) *is given satisfying* (4.2) *and* (4.3). *Then for all $\varepsilon > 0$ there exists a linear, time-invariant, finite-dimensional dynamic compensator F such that the closed-loop system is internally stable and has H_∞ norm less than ε.*

Remark. We note that even if for this new system we can make the H_∞ norm arbitrarily small, for the original system we are only sure that the H_∞ norm will be less than one. It is very well possible that a compensator for the new system yields an H_∞ norm of say 0.0001 while the same compensator makes the H_∞ norm of the original plant only 0.9999.

Before we can prove this result we have to do some preparatory work. We first have to introduce a number of subspaces from geometric control theory as follows.

DEFINITION 4.2. Assume we have a system

$$\Sigma_{ci}: \begin{cases} \dot{x} = Ax + Bu, \\ y = C_2x + D_2u. \end{cases} \tag{4.4}$$

We define *the strongly controllable subspace* $\mathcal{T}(\Sigma_{ci})$ as the smallest subspace $\mathcal{T}$ of $\mathcal{R}^n$ for which there exists a mapping G such that

$$(A + GC_2)\mathcal{T} \subset \mathcal{T}, \tag{4.5}$$

$$\operatorname{Im}(B + GD_2) \subset \mathcal{T}. \tag{4.6}$$

We also define the subspace $\mathcal{T}_g(\Sigma_{ci})$ as the smallest subspace $\mathcal{T}$ of $\mathcal{R}^n$ for which there exists a matrix G such that (4.5) and (4.6) are satisfied and, moreover, $A + GC_2 | \mathcal{R}^n / \mathcal{T}$ is asymptotically stable. It is well known that these subspaces are well defined in this way. A system is called *strongly controllable* if its strongly controllable subspace is equal to the whole state space.

We also define the dual versions of these subspaces as follows.

DEFINITION 4.3. Assume we have a system

$$\Sigma_{di}: \begin{cases} \dot{x} = Ax + Ew, \\ y = C_1x + D_1u. \end{cases} \tag{4.7}$$

We define *the weakly unobservable subspace* $\mathcal{V}(\Sigma_{di})$ as the largest subspace $\mathcal{V}$ of $\mathcal{R}^n$ for which there exists a mapping F such that

$$(A + EF)\mathcal{V} \subset \mathcal{V}, \tag{4.8}$$

$$(C_1 + D_1F)\mathcal{V} = \{0\}. \tag{4.9}$$

We also define the subspace $\mathcal{V}_g(\Sigma_{di})$ as the largest subspace $\mathcal{V}$ for which there exists a mapping F such that (4.8) and (4.9) are satisfied and, moreover, $A + EF | \mathcal{V}$ is asymptotically stable. It is well known that these subspaces are well defined in this way. A system is called *strongly observable* if its weakly unobservable subspace is equal to $\{0\}$.

In order to calculate these subspaces the following lemma will come in handy.

LEMMA 4.4. *$\mathcal{T}(\Sigma_{ci})$ equals the limit of the following sequence of subspaces*:

$$\begin{aligned} &\mathcal{T}_0(\Sigma_{ci}) := 0, \quad \mathcal{T}_{i+1}(\Sigma_{ci}) := \{x \in \mathcal{R}^n \mid \exists \tilde{x} \in \mathcal{T}_i(\Sigma_{ci}),\ u \in \mathcal{R}^m \textit{ such that} \\ &\qquad\qquad x = A\tilde{x} + Bu \textit{ and } C_2\tilde{x} + D_2u = 0\}. \end{aligned} \tag{4.10}$$

It is well known (see [16]) *that* $\mathcal{T}_i(\Sigma_{ci})$ $(i=1, 2, \cdots)$ *is a nondecreasing sequence of subspaces that attains its limit in a finite number of steps. In the same way* $\mathcal{V}(\Sigma_{di})$ *equals the limit of the following sequence of subspaces*:

$$\mathcal{V}_0(\Sigma_{di}) := \mathcal{R}^n, \quad \mathcal{V}_{i+1}(\Sigma_{di}) := \{x \in \mathcal{R}^n \mid \exists \tilde{u} \in \mathcal{R}^m, \text{ such that } Ax + E\tilde{u} \in \mathcal{V}_i(\Sigma_{di}) \text{ and } C_1 x + D_1 \tilde{u} = 0\}. \tag{4.11}$$

Moreover, if G is a mapping such that (4.5) *and* (4.6) *are satisfied for* $\mathcal{T} = \mathcal{T}(\Sigma_{ci})$ *and if F is a mapping such that* (4.8) *and* (4.9) *are satisfied for* $\mathcal{V} = \mathcal{V}(\Sigma_{di})$, *then we have the following two equalities*:

$$\mathcal{T}_g(\Sigma_{ci}) = [\mathcal{T}(\Sigma_{ci}) + \mathcal{X}_b(A+GC_2)] \cap \langle \mathcal{T}(\Sigma_{ci}) + C_2^{-1} \operatorname{im} D_2 \mid A + GC_2 \rangle, \tag{4.12}$$

$$\mathcal{V}_g(\Sigma_{di}) = \mathcal{V}(\Sigma_{di}) \cap \mathcal{X}_g(A+EF) + < A+EF \mid \mathcal{V}(\Sigma_{di}) \cap E \ker D_1 \rangle. \tag{4.13}$$

Here $\mathcal{X}_b(A+GC_2)$ *denotes the modal subspace of the matrix* $A+GC_2$ *with respect to the closed right halfplane and* $\mathcal{X}_g(A+EF)$ *denotes the modal subspace of the matrix* $A+EF$ *with respect to the open left halfplane. Finally,* $\langle A+EF \mid \mathcal{V}(\Sigma_{di}) \cap E \ker D_1 \rangle$ *denotes the smallest* $A+EF$ *invariant subspace containing* $\mathcal{V}(\Sigma_{di}) \cap E \ker D_1$ *and* $\langle \mathcal{T}(\Sigma_{ci}) + C_2^{-1} \operatorname{im} D_2 \mid A + GC_2 \rangle$ *denotes the largest* $A+GC_2$ *invariant subspace contained in* $\mathcal{T}(\Sigma_{ci}) + C_2^{-1} \operatorname{im} D_2$.

Proof. The proof is almost entirely well known except possibly (4.12) and (4.13) in case the D-matrices are unequal to zero. This can be proven by first showing that there exists a G satisfying (4.5) and (4.6) for which (4.12) holds and after that, showing that the equality is independent of our particular choice of G satisfying (4.5) and (4.6). The same can be done for (4.13). Details are left to the reader. □

We can express the rank conditions (4.2) and (4.3) in terms of these subspaces (see [3], [17]) as follows.

LEMMA 4.5. *Let system* (4.1) *be given. The rank condition* (4.2) *is satisfied if and only if*

$$\mathcal{V}_g(\Sigma_{ci}) + \mathcal{T}(\Sigma_{ci}) = \mathcal{R}^n. \tag{4.14}$$

The rank condition (4.3) *is satisfied if and only if*

$$\mathcal{V}(\Sigma_{di}) \cap \mathcal{T}_g(\Sigma_{di}) = \{0\}. \tag{4.15}$$

Here Σ_{ci} *is given by* (4.4) *and* Σ_{di} *is given by* (4.7).

Using this we can derive the following lemma.

LEMMA 4.6. *Let system* (4.1) *be given satisfying* (4.2) *and* (4.3). *For all* $\varepsilon > 0$ *there exist mappings F and G such that* $A+BF$ *and* $A+GC_1$ *are asymptotically stable and, moreover,*

$$\|(C_2 + D_2 F)(sI - A - BF)^{-1}\|_\infty < \varepsilon \tag{4.16}$$

and

$$\|(sI - A - GC_1)(E + GD_1)\|_\infty < \varepsilon. \tag{4.17}$$

Proof. By Definition 4.3 we know there exists a mapping $\tilde{F}$ such that

$$(A + B\tilde{F})\mathcal{V}_g(\Sigma_{ci}) \subset \mathcal{V}_g(\Sigma_{ci}), \tag{4.18}$$

$$(C_2 + D_2\tilde{F})\mathcal{V}_g(\Sigma_{ci}) = \{0\}, \tag{4.19}$$

and moreover, $A + B\tilde{F} \mid \mathcal{V}_g(\Sigma_{ci})$ is asymptotically stable. Define the canonical projection $\Pi : \mathcal{R}^n \to \mathcal{R}^n / \mathcal{V}_g(\Sigma)$. By (4.19) there exists a mapping $\bar{C}$ such that $C_2 + D_2\tilde{F} = \bar{C}\Pi$.

Moreover, by (4.18) there exists a mapping $\bar{A}$ such that $\Pi(A+B\tilde{F})=\bar{A}\Pi$. Finally, define $\bar{B}:=\Pi B$ and the system:

$$(4.20) \qquad \Sigma_{fs}:\begin{cases} \dot{p}=\bar{A}p+\bar{B}u, \\ z=\bar{C}p+D_2u. \end{cases}$$

It can be easily shown by induction using the algorithm (4.10) that $\mathcal{T}_i(\Sigma_{fs})=\Pi\mathcal{T}_i(\Sigma_{ci})$ for $i=0, 1, \cdots$. Hence we have

$$\mathcal{T}(\Sigma_{fs})=\Pi\mathcal{T}(\Sigma_{ci})=\Pi\{\mathcal{T}(\Sigma_{ci})+\mathcal{V}_g(\Sigma_{ci})\}=\Pi\mathcal{R}^n=\mathcal{R}^n/\mathcal{V}_g(\Sigma_{ci}).$$

This implies that the system (4.20) is strongly controllable.

Define F_0 such that $(\bar{C}+D_2F_0)^TD_2=0$ and define M such that $\ker D_2=\operatorname{im} M$. It can be easily checked that $\mathcal{T}(\Sigma_{fs})=\mathcal{T}(\bar{A}+\bar{B}F_0, \bar{B}M, \bar{C}+D_2F_0, 0)$. Hence by Theorem 3.36 of [21] we know there exist an $\bar{F}$ such that

$$(4.21) \qquad \|(\bar{C}+D_2F_0)\, e^{((\bar{A}+\bar{B}F_0)+\bar{B}M\bar{F})t}\|_1<\varepsilon$$

and such that $\bar{A}+\bar{B}F_0+\bar{B}M\bar{F}$ is asymptotically stable.

Define $F:=\tilde{F}+(F_0+M\bar{F})\Pi$; then

$$(4.22) \qquad A+BF\,|\,\mathcal{V}_g(\Sigma_{ci})=A+B\tilde{F}\,|\,\mathcal{V}_g(\Sigma_{ci}),$$

$$(4.23) \qquad \Pi(A+BF)=(\bar{A}+\bar{B}F_0+\bar{B}M\bar{F})\Pi.$$

It can be easily shown that this implies that $A+BF$ is asymptotically stable. Moreover, we have

$$(4.24) \qquad (C_2+D_2F)\, e^{(A+BF)t}=(\bar{C}+D_2F_0)\, e^{((\bar{A}+\bar{B}F_0)+\bar{B}M\bar{F})t}$$

for all $t>0$. Using (4.24), we find for all $s\in i\mathcal{R}$ (use that $|e^{st}|=1$):

$$\begin{aligned}\|(C_2+D_2F)(sI-(A+BF))^{-1}\| &= \left\|\int_0^\infty (C_2+D_2F)\, e^{((A+BF)-sI)t}\,dt\right\| \\ &\leqq \int_0^\infty \|(C_2+D_2F)\, e^{((A+BF)-sI)t}\|\,dt \\ &= \|(C_2+D_2F)\, e^{(A+BF)t}\|_1 \\ &= \|(\bar{C}+D_2F_0)\, e^{((\bar{A}+\bar{B}F_0)+\bar{B}M\bar{F})t}\|_1 \\ &\leqq \varepsilon.\end{aligned}$$

This implies (4.16). Therefore F satisfies all the requirements of the lemma. The existence of a G such that $A+GC$ is asymptotically stable and such that (4.17) is satisfied can be obtained by dualization. □

We can now prove Theorem 4.1.

Proof of Theorem 4.1. Let $\varepsilon>0$. We first choose a mapping F such that

$$(4.25) \qquad \|(C_2+D_2F)(sI-A-BF)^{-1}\|_\infty<\varepsilon/3\|E\|^{-1}$$

and such that $A+BF$ is asymptotically stable. This can be done according to Lemma 4.6. Next choose a mapping G such that

$$(4.26) \qquad \|(sI-A-GC_1)^{-1}(E+GD_1)\|_\infty<\min\{\varepsilon/3\|D_2F\|^{-1}, \|E\|\,\|BF\|^{-1}\}$$

and such that $A+GC_1$ is asymptotically stable. Again Lemma 4.6 guarantees the existence of such a G. We apply the following dynamic feedback compensator to the system (4.1):

$$\Sigma_{F,G}:\begin{cases}\dot{p}=Ap+Bu+G(C_1p-y),\\ u=Fp.\end{cases} \tag{4.27}$$

The closed-loop system is given by (where $e:=x-p$):

$$\Sigma_{\mathrm{cl}}:\begin{cases}\begin{pmatrix}\dot{x}\\ \dot{e}\end{pmatrix}=\begin{pmatrix}A+BF & -BF\\ 0 & A+GC_1\end{pmatrix}\begin{pmatrix}x\\ e\end{pmatrix}+\begin{pmatrix}E\\ E+GD_1\end{pmatrix}w,\\ z=(C_2+D_2F\quad -D_2F)\begin{pmatrix}x\\ e\end{pmatrix}.\end{cases} \tag{4.28}$$

It is clear that this is an internally stabilizing feedback. We now calculate the transfer matrix from w to z of this system:

$$\begin{aligned}&(C_2+D_2F)(sI-A-BF)^{-1}E\\ &\quad-(C_2+D_2F)(sI-A-BF)^{-1}BF(sI-A-GC_1)^{-1}(E+GD_1)\\ &\quad-D_2F(sI-A-GC_1)^{-1}(E+GD_1).\end{aligned}$$

Using (4.25) and (4.26) it can easily be shown that this closed-loop transfer matrix has H_∞ norm less than ε. □

We are now able to complete the proof of Theorem 2.1.

Proof of the implication (ii)⇒(i) *of Theorem* 2.1. Since we can transform the original system into a system satisfying (4.2) and (4.3) we know by Lemma 4.1 that we can find an internally stabilizing dynamic compensator for this new system which is such that the closed-loop transfer matrix has H_∞ norm less than one. By applying Corollary 3.4 we know that this compensator F satisfies the requirements in Theorem 2.1(i). □

5. The design of an admissible compensator. In this section we shall give a method to calculate a dynamic compensator F such that the closed-loop system is internally stable and, moreover, the closed-loop transfer matrix has H_∞ norm less than one. We shall derive this F step by step, using the following conceptual algorithm.

(i) Calculate P and Q satisfying part (ii) of Theorem 2.1. This can, for instance, be done using Lemma A4. If they do not exist or if $\rho(PQ)\geqq 1$, then there does not exist a dynamic feedback satisfying part (i) of Theorem 2.1 and we stop.

(ii) Perform the factorizations (3.1) and (3.11). We can now construct the system $\Sigma_{P,Q}$ as given by (3.12).

We now start solving the almost disturbance decoupling problem for the system (3.12) we obtained in step (ii). As in § 4 we shall rename our variables and assume that we have a system in the form (4.1). We set $\varepsilon=1$. We have to construct matrices F and G such that (4.25) and (4.26) are satisfied and, moreover, such that $A+BF$ and $A+GC_1$ are asymptotically stable. We shall only discuss the construction of F. The construction of G can be obtained by dualization.

(iii) Construct $\mathcal{V}_g(\Sigma_{ci})$ by using Lemma 4.4.

(iv) Construct an $\tilde{F}$ such that (4.18) and (4.19) are satisfied and, moreover, such that $A+B\tilde{F}\,|\,\mathcal{V}_g(\Sigma_{ci})$ is asymptotically stable.

(v) Define the canonical projection $\Pi:\mathscr{R}^n \to \mathscr{R}^n/\mathscr{V}_g(\Sigma)$ and the mappings $\bar{A}$, $\bar{B}$, and $\bar{C}$ satisfying:
   (1) $\Pi(A+B\tilde{F})=\bar{A}\Pi$,
   (2) $\bar{B}:=\Pi B$,
   (3) $C_2+D_2\tilde{F}=\bar{C}\Pi$.
   Construct the system Σ_{fs} as given by (4.20).

(vi) Construct F_0 such that $(\bar{C}+D_2F_0)^T D_2=0$ and M such that $\operatorname{im}\bar{B}M=\bar{B}\ker D_2$. Define the following matrices:
   (1) $\tilde{A}:=\bar{A}+\bar{B}F_0$,
   (2) $\tilde{B}:=\bar{B}M$,
   (3) $\tilde{C}:=\bar{C}+D_2F_0$,
   and the system

$$\Sigma_{ht}:\begin{cases}\dot{x}=\tilde{A}x+\tilde{B}u,\\ z=\tilde{C}x.\end{cases} \tag{5.1}$$

In this way we obtained a strongly controllable system (5.1), for which we have to find a static feedback $\bar{F}$ such that the closed-loop system is internally stable and such that the closed-loop impulse response satisfies the $\mathscr{L}_1$ norm bound $\varepsilon/3\|E\|^{-1}$. We shall use a method for this which was given in [21].

(vii) We construct a new basis for the state space. We shall construct it by induction. Choose $x_1\in\ker\tilde{C}\cap\operatorname{im}\tilde{B}$ and v_1 such that $x_1=\tilde{B}v_1$. If x_1 does not exist go to item (viii). Assume $\{x_1,\cdots,x_i\}$ and $\{v_1,\cdots,v_i\}$ are given. Denote by $\mathscr{S}_i$ the linear span of $\{x_1,\cdots,x_i\}$. If $\{\tilde{A}x_i+\operatorname{im}\tilde{B})\cap\ker\tilde{C}\subset\mathscr{S}_i$ and $\operatorname{im}\tilde{B}\cap\ker\tilde{C}\subset\mathscr{S}_i$, then goto step (viii). Otherwise, if $(\tilde{A}x_i+\operatorname{im}\tilde{B})\cap\ker\tilde{C}\not\subset\mathscr{S}_i$, then choose v such that $\tilde{A}x_i+\tilde{B}v\in\ker\tilde{C}$ and $\tilde{A}x_i+\tilde{B}v\notin\mathscr{S}_i$. Set $x_{i+1}=\tilde{A}x_i+\tilde{B}v$ and $v_{i+1}=v$. (If $(\tilde{A}x_i+\operatorname{im}\tilde{B})\cap\ker\tilde{C}\subset\mathscr{S}_i$, then choose v such that $\tilde{B}v\in\ker\tilde{C}$ and $\tilde{B}v\notin\mathscr{S}_i$. Set $x_{i+1}=\tilde{B}v$ and $v_{i+1}=v$. Set $i:=i+1$ and repeat this paragraph again.

(viii) Define $\mathscr{R}_a^*(\ker\tilde{C})=\mathscr{S}_i$. Define a linear mapping F such that $Fx_j=v_j$, $j=1,\cdots,i$ and extend it to the whole state space. In [21] it has been shown that $A\mathscr{R}_a^*(\ker\tilde{C})+\operatorname{im}\tilde{B}=\mathscr{T}(\Sigma_{ht})=\mathscr{R}^n$. Therefore it is easily seen that we can extend $\{x_1,\cdots,x_i\}$ to a basis of $\mathscr{R}^n$ which can be written as

$$\begin{matrix}\tilde{B}v_1, & A_F\tilde{B}v_1, & \cdots, & A_F^{r_1}\tilde{B}v_1,\\ \tilde{B}v_2, & A_F\tilde{B}v_2, & \cdots, & A_F^{r_2}\tilde{B}v_2,\\ \vdots & \vdots & & \vdots\\ \tilde{B}v_j, & A_F\tilde{B}v_j, & \cdots, & A_F^{r_j}\tilde{B}v_j,\end{matrix}$$

where $A_F=\tilde{A}+\tilde{B}F$ and for those $k=1,\cdots,j$ for which $r_k\geqq 1$ we have $\tilde{B}v_k$, $A_F\tilde{B}v_k,\cdots,A_F^{r_k-1}\tilde{B}v_k\in\ker\tilde{C}$.

(ix) We define the following sequence of vectors. For $i=1,\cdots,j$ we define:

$$\begin{aligned} x_{i,1}(n) &:= \left(I+\frac{1}{n}A_F\right)^{-1}\tilde{B}v_i\\ x_{i,2}(n) &:= \left(I+\frac{1}{n}A_F\right)^{-1}A_F x_{i,1}(n)\\ \vdots\quad & \quad\vdots\\ x_{i,r_i+1}(n) &:= \left(I+\frac{1}{n}A_F\right)^{-1}A_F x_{i,r_i}(n). \end{aligned}$$

Since $x_{i,k}(n) \to A_F^{k-1}\tilde{B}v_i$ as $n \to \infty$ for $i = 1, \cdots, j$ and $k = 1, \cdots, r_i + 1$ it can be easily seen that for n sufficiently large the vectors $\{x_{i,k}(n),\ i = 1, \cdots, j;\ k = 1, \cdots, r_i + 1\}$ are linearly independent and hence form a basis of $\mathcal{R}^n$ again. Let N be such that for all $n > N$ these vectors indeed form a basis.

(x) For all $n > N$ define a linear mapping $\bar{F}_n$ by

$$\begin{aligned} \bar{F}_n x_{i,1}(n) \quad &:= -nv_i \\ \bar{F}_n x_{i,2}(n) \quad &:= -n^2 v_i \\ \vdots \qquad & \qquad \vdots \\ \bar{F}_n x_{i,r_i+1}(n) \quad &:= -n^{r_i+1} v_i. \end{aligned}$$

This determines $\bar{F}_n$ uniquely. Define $F_n := F + \bar{F}_n$. It is shown in [21] that the spectrum of $\tilde{A} + \tilde{B}F_n$ is the set $\{-n\}$. Moreover, we have

$$\lim_{n\to\infty} \|\tilde{C}\, e^{(A+BF_n)t}\|_1 = 0.$$

Choose n such that the impulse response satisfies the required $\mathcal{L}_1$ bound $\varepsilon/3\|E\|^{-1}$. This F_n is internally stabilizing and satisfies the $\mathcal{L}_1$ bound. Now we can construct the F we were looking for:

(xi) Define $F = \tilde{F} + (F_0 + MF_n)\Pi$. This F is internally stabilizing and is such that (4.25) is satisfied.

We construct G by dualizing the construction of F and the required dynamic compensator is finally given by (4.27).

6. Conclusion. In this paper we have given a complete treatment of the H_∞ problem with measurement feedback without restrictions on the direct feedthrough matrices. It remains however an open problem how we can treat invariant zeros on the imaginary axis. Other open problems are the minimally required dynamic order of the controller and the behaviour of the feedbacks and closed-loop system if we make the bound γ tighter. The latter problem has been investigated previously. It is possible that the infimum can only be attained by a nonproper controller (see [4]). But using the ideas of this paper it is perhaps possible to characterize whether or not this problem arises.

Finally, it would be interesting to characterize all solutions. In our opinion it is, however, in general not possible to obtain a characterization similar to the one obtained in [2]. This is due to the fact that the so-called central controller can be nonproper.

In our opinion this paper gives support to our claim that the approach to solve the H_∞ problem in the time-domain is a much more intuitive and appealing approach than the other methods used in recent papers.

Appendix A. A preliminary system transformation. In this section we shall choose bases in input, output, and state space that will give us much more insight into the structure of our problem. Although these decompositions are not necessary in the formulation of the main steps of the proof of Theorem 2.1, the details of the proof are very much concerned with these decompositions. It will be shown that the matrices defining our systems in these bases have a very particular structure. For details we refer to [18]. We shall display this structure by writing down the matrices with respect to these suitably chosen bases for the input, state, and output spaces.

Our basic tool is the strongly controllable subspace. This subspace has already been defined in Definition 4.2.

We shall give one property of the strongly controllable subspace at this point which will come in handy in the sequel (see [7], [16]).

LEMMA A1. *Consider the system* (4.4). *The system is strongly controllable if and only if*

$$\text{(A1)} \qquad \begin{pmatrix} sI-A & -B \\ C_2 & D_2 \end{pmatrix}$$

has rank $n+\operatorname{rank}(C_2 \quad D_2)$ *for all* $s \in \mathscr{C}$.

We can now define the bases for the system (2.1) which will be used in the sequel. It is also possible to define a dual version of this decomposition but we will only need this one. First choose a basis of the control input space $\mathscr{R}^m$. Decompose $\mathscr{R}^m = \mathscr{U}_1 \oplus \mathscr{U}_2$ such that $\mathscr{U}_2 = \ker D_2$ and $\mathscr{U}_1$ arbitrary. Choose a basis $u_1, u_2, \cdots, u_m$ of $\mathscr{R}^m$ such that $u_1, u_2, \cdots, u_i$ is a basis of $\mathscr{U}_1$ and $u_{i+1}, \cdots, u_m$ is a basis of $\mathscr{U}_2$.

Next choose *an orthonormal basis* $z_1, z_2, \cdots, z_p$ of the output space $\mathscr{R}^p$ such that $z_1, \cdots, z_j$ is a basis of im D_2 and $z_{j+1}, \cdots, z_p$ is a basis of $(\text{im } D_2)^\perp$. Because this is an orthonormal basis this basis transformation does not change the norm $\|z\|$.

Finally, we choose a decomposition of the state space $\mathscr{R}^n = \mathscr{X}_1 \oplus \mathscr{X}_2 \oplus \mathscr{X}_3$ such that $\mathscr{X}_2 = \mathscr{T}(\Sigma_{ci}) \cap C_2^{-1} \text{ im } D_2$, $\mathscr{X}_2 \oplus \mathscr{X}_3 = \mathscr{T}(\Sigma_{ci})$ and $\mathscr{X}_1$ arbitrary. We choose a corresponding basis $x_1, x_2, \cdots, x_n$ such that $x_1, \cdots, x_r$ is a basis of $\mathscr{X}_1$, $x_{r+1}, \cdots, x_s$ is a basis of $\mathscr{X}_2$ and $x_{s+1}, \cdots, x_n$ is a basis of $\mathscr{X}_3$.

With respect to these bases the maps B, C_2, and D_2 have the following form:

$$\text{(A2)} \qquad B = (B_1 \quad B_2), \quad C_2 = \begin{pmatrix} \hat{C}_1 \\ \hat{C}_2 \end{pmatrix}, \quad D_2 = \begin{pmatrix} \hat{D}_2 & 0 \\ 0 & 0 \end{pmatrix},$$

where $\hat{D}_2$ is invertible. Next, we define a linear mapping $F_0 : \mathscr{R}^n \to \mathscr{R}^m$ by

$$\text{(A3)} \qquad F_0 := \begin{pmatrix} -\hat{D}_2^{-1}\hat{C}_1 \\ 0 \end{pmatrix} \quad \text{and hence } C_2 + D_2F_0 = \begin{pmatrix} 0 \\ \hat{C}_2 \end{pmatrix}.$$

We have the following properties of this decomposition which are proven in [18].

LEMMA A2. *Let* F_0 *be given by* (A3). *Then we have*

(i) $(A+BF_0)(\mathscr{T}(\Sigma) \cap C_2^{-1} \text{ im } D_2) \subseteq \mathscr{T}(\Sigma)$,

(ii) im $B_2 \subseteq \mathscr{T}(\Sigma)$,

(iii) $\mathscr{T}(\Sigma) \cap C_2^{-1} \text{ im } D_2 \subseteq \ker \hat{C}_2$.

By applying this lemma we find that the matrices $A+BF_0$, B, $C_2+D_2F_0$ and D_2 with respect to these bases have the following form:

$$\text{(A4)} \qquad \begin{aligned} A+BF_0 &= \begin{pmatrix} A_{11} & 0 & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}, & B &= \begin{pmatrix} B_{11} & 0 \\ B_{21} & B_{22} \\ B_{31} & B_{32} \end{pmatrix}, \\ C_2+D_2F_0 &= \begin{pmatrix} 0 & 0 & 0 \\ C_{21} & 0 & C_{23} \end{pmatrix}, & D_2 &= \begin{pmatrix} \hat{D}_2 & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

We decompose the matrices $\hat{C}_1$ and E correspondingly:

$$\text{(A5)} \qquad \hat{C}_1 = (C_{11} \quad C_{12} \quad C_{13}), \qquad E = \begin{pmatrix} E_1 \\ E_2 \\ E_3 \end{pmatrix}.$$

These matrices turn out to have some nice structural properties, which have been shown in [18].

LEMMA A3. *We have the following properties*:

(i) C_{23} *is injective,*

(ii) *the system*

$$\Sigma_1 := \left[\begin{pmatrix} A_{22} & A_{23} \\ A_{32} & A_{33} \end{pmatrix}, \begin{pmatrix} B_{22} \\ B_{32} \end{pmatrix}, (0 \quad I), 0\right] \tag{A6}$$

is strongly controllable,

(iii) *we have*

$$\text{normrank } G = \text{rank} \begin{pmatrix} C_{23} & 0 \\ 0 & \hat{D}_2 \end{pmatrix}, \tag{A7}$$

where G is the transfer matrix defined by $G(s) := C_2(sI-A)^{-1}B + D_2$.

We need the following results from [18] which connects the conditions of Theorem 2.1 to the matrices as defined in [A4].

LEMMA A4. *Assume* $P \in \mathcal{R}^{n\times n}$ *is symmetric and* $F(P) \geqq 0$. *Then we have the following*:

(i) $P\mathcal{T}(\Sigma) = 0$, *i.e., in our decomposition P can be written as*

$$P = \begin{pmatrix} P_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{A8}$$

(ii) *If P has the form* (A8), *then*

$$\begin{aligned} R(P_1) &:= P_1A_{11} + A_{11}^TP_1 + C_{21}^TC_{21} + P_1(E_1E_1^T - B_{11}(\hat{D}_2^T\hat{D}_2)^{-1}B_{11}^T)P_1 \\ &\quad -(P_1A_{13} + C_{21}^TC_{23})(C_{23}^TC_{23})^{-1}(A_{13}^TP_1 + C_{23}^TC_{21}) \geqq 0. \end{aligned} \tag{A9}$$

Moreover, $R(P_1) = 0$ *if and only if* rank $F(P) =$ normrank G.

(iii) *If* $R(P_1) = 0$, *then we have*

$$\text{rank} \begin{pmatrix} L(P,s) \\ F(P) \end{pmatrix} = n + \text{normrank } G \quad \forall s \in \mathcal{C}^0 \cup \mathcal{C}^+$$

if and only if

$$Z(P_1) := A_{11} + E_1E_1^TP_1 - B_{11}(\hat{D}_2^T\hat{D}_2)^{-1}B_{11}^TP_1 - A_{13}(C_{23}^TC_{23})^{-1}(A_{13}^TP_1 + C_{23}^TC_{21})$$

is an asymptotically stable matrix. Moreover, in that case also the matrix

$$A_{11} - B_{11}(\hat{D}_2^T\hat{D}_2)^{-1}B_{11}^TP_1 - A_{13}(C_{23}^TC_{23})^{-1}(A_{13}^TP_1 + C_{23}^TC_{21})$$

is an asymptotically stable matrix.

COROLLARY A5. *If there exists a matrix* $P \geqq 0$ *such that* $F(P) \geqq 0$ *and moreover*:

(i) rank $F(P) =$ normrank G,

(ii) $\text{rank} \begin{pmatrix} L(P,s) \\ F(P) \end{pmatrix} = n + \text{normrank } G \quad \forall s \in \mathcal{C}^0 \cup \mathcal{C}^+$,

then this matrix is uniquely defined by the above inequality and the corresponding two rank conditions.

Proof. By Lemma A4 a solution P must be of the form (A8) where P_1 is a solution of the algebraic Riccati equation $R(P_1) = 0$ such that $Z(P_1)$ is asymptotically stable. Denote the Hamiltonian matrix corresponding to this algebraic Riccati equation by

H; then we have

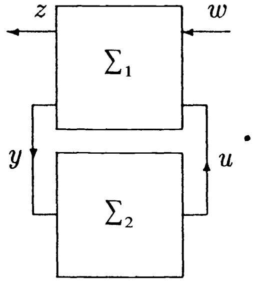
$$H\begin{pmatrix} I \\ P_1 \end{pmatrix} = \begin{pmatrix} I \\ P_1 \end{pmatrix} Z(P_1). \tag{A10}$$

Since a Hamiltonian matrix has the property that λ is an eigenvalue if and only if $-\lambda$ is an eigenvalue of H, we know that an n-dimensional invariant subspace $\mathcal{W}$ of H such that $H \mid \mathcal{W}$ is asymptotically stable must be unique. This implies that P_1 is unique and hence also P is unique. □

Appendix B. Proofs concerning the system transformations. In order to prove Lemma 3.2 we must first do some preparatory work. We first recall the following lemma from [2] which we shall use in the sequel.

LEMMA B1. *Suppose we have the following interconnection of two systems Σ_1 and Σ_2, both described by some state-space representation*:

(B1)



Assume Σ_1 is internally stable and its transfer matrix L from $\binom{w}{u}$ to $\binom{z}{y}$ satisfies $L^{\sim}L = I$ where $L^{\sim}(s) := L^T(-s)$. Moreover, assume that if we decompose L:

$$L =: \begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} \tag{B2}$$

compatible with the sizes of w, u, z, and y, we have $L_{21}^{-1} \in H_\infty$ and $\lim_{s\to\infty} L_{22}(s) = 0$. Then the following two statements are equivalent:

(i) *The closed-loop system* (B1) *is internally stable and its closed-loop transfer matrix has H_∞ norm less than one.*

(ii) *The system Σ_2 is internally stable and its transfer matrix has H_∞ norm less than one.*

Proof. This is a well-known result although written down here in a different way. Note that if the closed-loop system (B1) is internally stable, then Σ_2 is stabilizable and detectable. This can be shown either by writing down the closed-loop differential equation or by noting that an unstable uncontrollable mode in Σ_2 cannot be controlled by y and hence is still unstable and uncontrollable in the closed-loop system and the same for an unstable unobservable mode. The result in this form can then be obtained by using the work in [14]. □

We shall now assume that we have chosen the bases described in Appendix A. Let P satisfy the conditions of Lemma 3.1(i). Hence we know P has the form (A8). It is easily shown that it is sufficient to prove the lemma for one specific choice of $C_{2,P}$ and D_P. We define the following matrices:

$$C_{2,P} := \begin{pmatrix} \hat{D}_2(\hat{D}_2^T\hat{D}_2)^{-1}B_{11}^TP_1 + C_{11} & C_{12} & C_{13} \\ C_{23}(C_{23}^TC_{23})^{-1}(A_{13}^TP_1 + C_{23}^TC_{21}) & 0 & C_{23} \end{pmatrix}, \tag{B3}$$

(B4) $$D_P := \begin{pmatrix} \hat{D}_2 & 0 \\ 0 & 0 \end{pmatrix} (= D_2).$$

By writing down $F(P)$ in terms of the chosen bases and by using the fact that P_1 satisfies the algebraic Riccati equation $R(P_1) = 0$ where $R(P_1)$ is defined by (A9), it can be checked after some effort that these matrices indeed satisfy (3.1). We define the following matrices:

(B5) $$\tilde{A} := A_{11} - A_{13}(C_{23}^T C_{23})^{-1}(A_{13}^T P_1 + C_{23}^T C_{21}) - B_{11}(\hat{D}_2^T \hat{D}_2)^{-1} B_{11}^T P_1,$$

(B6) $$\tilde{C}_1 := -(\hat{D}_2^T)^{-1} B_{11}^T P_1,$$

(B7) $$\tilde{C}_2 := C_{21} - C_{23}(C_{23}^T C_{23})^{-1}(A_{13}^T P_1 + C_{23}^T C_{21}),$$

(B8) $$\tilde{B}_{11} := B_{11} \hat{D}_2^{-1},$$

(B9) $$\tilde{B}_{12} := A_{13}(C_{23}^T C_{23})^{-1} C_{23}^T - P_1^{\dagger} C_{21}^T (I - C_{23}(C_{23}^T C_{23})^{-1} C_{23}^T),$$

where † denotes the Moore–Penrose inverse. We now define the following system:

(B10) $$\Sigma_U : \begin{cases} \dot{x}_U = \tilde{A} x_U + (\tilde{B}_{11} \quad \tilde{B}_{12}) u + E_1 w, \\ y_U = -E_1^T P_1 x_U + w, \\ z_U = \begin{pmatrix} \tilde{C}_1 \\ \tilde{C}_2 \end{pmatrix} x_U + \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} u. \end{cases}$$

We have the following properties of the system Σ_U.

LEMMA B2. *The system Σ_U is internally stable. Let U denote the transfer matrix of Σ_U from $\binom{u}{w}$ to $\binom{y_u}{z_u}$. We have $U^{\sim} U = I$ where $U^{\sim}(s) := U^T(-s)$. If we decompose U:*

(B11) $$U := \begin{pmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{pmatrix}$$

compatible with the sizes of u, w, y_U, and z_U then we have $U_{21}^{-1} \in H_\infty$ and $\lim_{s \to \infty} U_{22}(s) = 0$.

Proof. The fact that Σ_U is internally stable and that $U_{21}^{-1} \in H_\infty$ follows directly from the fact that $\tilde{A}$ and $\tilde{A} + E_1 E_1^T P_1$ are asymptotically stable by Lemma A4(iii). The fact that $\lim_{s \to \infty} U_{22}(s) = 0$ can be checked trivially. It can be easily checked using Lemma A4(ii) that P_1 is the controllability gramian of Σ_U. Moreover, we have

(B12) $$\begin{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} \\ I & (0 \quad 0) \end{pmatrix} \begin{pmatrix} -E_1^T P_1 \\ \begin{pmatrix} \tilde{C}_1 \\ \tilde{C}_2 \end{pmatrix} \end{pmatrix} + \begin{pmatrix} \begin{pmatrix} \tilde{B}_{11}^T \\ \tilde{B}_{12}^T \end{pmatrix} \\ E_1^T \end{pmatrix} P_1 = 0.$$

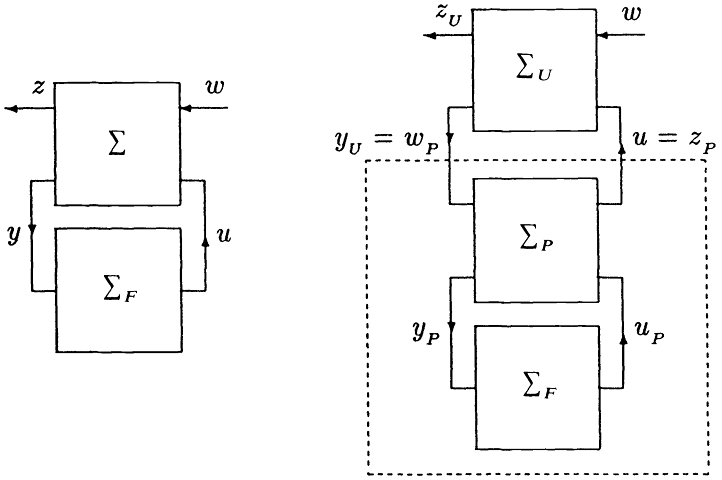
This can be checked by simply writing out and using the fact that

$$\ker P_1 \subset \ker (I - C_{23}(C_{23}^T C_{23})^{-1} C_{23}^T) C_{21}.$$

The result that $U \sim U = I$ then follows by applying Theorem 5.1 of [5]. □

Proof of Lemma 3.2. We have our special choice of $C_{2,P}$ and D_P given by (B3) and (B4). As we have already noted, taking this special choice for C_P and D_P is not essential. We shall first compare the following two systems:

(B13)



The system on the left is the same as the system on the left in (3.3), and the system on the right is described by the system (B10) interconnected with the system on the right in (3.3). We decompose the state of Σ, x into x_1, x_2, and x_3 according to the choice of bases described in Appendix A and decompose the state of Σ_P into $x_{1,P}$, $x_{2,P}$, $x_{3,P}$ of corresponding sizes. (Note that Σ and Σ_P have the same state space $\mathscr{R}^n$.) Writing out all the differential equations using the decompositions of the matrices given in (A3)-(A5) we find

$$\left\{\begin{array}{l} \begin{pmatrix} \dot{x}_U - \dot{x}_{1,P} \\ \dot{x}_P \\ \dot{p} \end{pmatrix} = \begin{pmatrix} \tilde{A} + E_1 E_1^T P_1 & 0 & 0 \\ * & A + BNC_1 & BM \\ * & LC_1 & K \end{pmatrix} \begin{pmatrix} x_U - x_{1,P} \\ x_P \\ p \end{pmatrix} + \begin{pmatrix} 0 \\ E + BND_1 \\ LD_1 \end{pmatrix} w, \\ \qquad z = (* \quad C_2 + D_2NC_1 \quad D_2M) \begin{pmatrix} x_U - x_{1,P} \\ x_P \\ p \end{pmatrix} + D_2ND_1\omega. \end{array}\right.$$

The $*$ denotes matrices which are unimportant for this argument. The system on the right is internally stable if and only if the system described by the above set of equations is internally stable. If we also derive the system equations for the system on the left in (B13) we immediately see that, since $\tilde{A} + E_1 E_1^T P_1$ is asymptotically stable, the system on the left is internally stable if and only if the system on the right is internally stable. Moreover, if we take zero initial conditions and both systems have the same input w, then we have $z = z_U$, i.e., the input-output behaviour of both systems are equivalent. Hence the system on the left has H_∞ norm less than one if and only if the system on the right has H_∞ norm less than one.

By Lemma B2 we may apply Lemma B1 to the system on the right in (B13) and hence we find that the closed-loop system is internally stable and has H_∞ norm less than one if and only if the dashed system is internally stable and has H_∞ norm less than one.

Since the dashed system is exactly the system on the right in (3.3) and the system on the left in (B13) is exactly equal to the system on the left in (3.3), we have completed the proof. □

We will now prove Lemma 3.3. In fact, we will prove the dual version of this lemma since this is much more convenient to us. We first factorize $G(Q)$:

$$G(Q) := \begin{pmatrix} E_Q \\ D_Q \end{pmatrix} (E_Q^T \quad D_Q^T). \tag{B14}$$

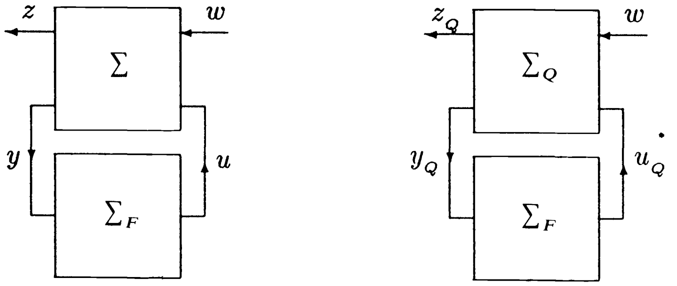
Define $A_Q := A + QC_2^TC_2$ and $B_Q := B + QC_2^TD_2$ and the system:

$$\Sigma_Q : \begin{cases} \dot{x}_Q = A_Qx_Q + B_Qu_Q + E_Qw, \\ y_Q = C_1x_Q \qquad\qquad + D_Qw \\ z_Q = C_2x_Q + D_2u_Q. \end{cases} \tag{B15}$$

By using the well-known facts that F stabilizes Σ if and only if F^T stabilizes Σ^T and $\|G\|_\infty = \|G^T\|_\infty$, we can derive the following dualized version of Lemma 3.2 for this dual system as follows.

LEMMA B3. *Let Q satisfy Lemma* 3.1(ii). *Moreover, let an arbitrary linear time-invariant finite-dimensional compensator F be given, described by* (2.2). *Let the following two systems be given where the system on the left is the interconnection of* (2.1) *and* (2.2) *and the system on the right is the interconnection of* (B15) *and* (2.2).

(B16)



Then the following statements are equivalent:

(i) *The system on the left is internally stable and its transfer matrix has H_∞ norm less than one.*

(ii) *The system on the right is internally stable and its transfer matrix has H_∞ norm less than one.*

We now investigate how the matrices appearing in the matrix inequality and the rank conditions look like for this new system Σ_Q:

$$\tilde{F}(X) := \begin{pmatrix} A_Q^TX + XA_Q + C_2^TC_2 + XE_QE_Q^TX & XB_Q + C_2^TD_2 \\ B_Q^TX + D_2^TC_2 & D_2^TD_2 \end{pmatrix}, \tag{B17}$$

$$\tilde{G}(Y) := \begin{pmatrix} A_QY + YA_Q^T + E_QE_Q^T + YC_2^TC_2Y & YC_1^T + E_QD_Q^T \\ C_1Y + D_QE_Q^T & D_QD_Q^T \end{pmatrix}, \tag{B18}$$

$$\tilde{L}(X, s) := (sI - A_Q - E_QE_Q^TX \quad -B_Q), \tag{B19}$$

$$\tilde{M}(Y, s) := \begin{pmatrix} sI - A_Q - YC_2^TC_2 \\ -C_1 \end{pmatrix}. \tag{B20}$$

Moreover, we define two new transfer matrices:

$$\tilde{G}(s) := C_2(sI - A_Q)^{-1}B_Q + D_2, \tag{B21}$$

$$\tilde{H}(s) := C_1(sI - A_Q)^{-1}E_Q + D_Q. \tag{B22}$$

Using these definitions we have the following result.

LEMMA B4. *Let Q satisfy Lemma* 3.1(ii). *Then $Y=0$ is the unique solution of the quadratic matrix inequality $G(Y)\geqq 0$ satisfying the following rank conditions*:

(1) $\operatorname{rank}\tilde{G}(Y)=\operatorname{normrank}\tilde{H}$,

(2) $\operatorname{rank}(\tilde{M}(Y,s)\quad \tilde{G}(Y))=n+\operatorname{normrank}\tilde{H}\quad \forall s\in\mathscr{C}^0\cup\mathscr{C}^+$.

Proof. It is trivial to check that $\tilde{G}(0)\geqq 0$. Moreover, since $\tilde{G}(0)=G(Q)$ and $\tilde{M}(0,s)=M(Q,s)$ it remains to show that normrank $\tilde{H}=$ normrank H. We have

$$\begin{aligned}\operatorname{normrank}\tilde{H}&=\operatorname{normrank}\begin{pmatrix} sI-A_Q & E_Q\\ -C_1 & D_Q\end{pmatrix}-n\\ &=\operatorname{normrank}\begin{pmatrix} sI-A_Q & E_QE_Q^T & D_QE_Q^T\\ -C_1 & D_QE_Q^T & D_QD_Q^T\end{pmatrix}-n\\ &=\operatorname{normrank}(M(Q,s)\quad G(Q))-n\\ &=\operatorname{normrank}H.\end{aligned}$$

Y is unique by Corollary A5. This is exactly what we had to prove. □

LEMMA B5. *There exists a solution X of the matrix inequality $\tilde{F}(X)\geqq 0$ satisfying the following two rank conditions*:

(1) $\operatorname{rank}\tilde{F}(X)=\operatorname{normrank}\tilde{G}$,

(2) $\operatorname{rank}\begin{pmatrix} r\tilde{L}(X,s)\\ \tilde{F}(X)\end{pmatrix}=n+\operatorname{normrank}\tilde{G}\quad \forall s\in\mathscr{C}^0\cup\mathscr{C}^+$,

if and only if $I-PQ$ is invertible. Moreover, in that case the solution is unique and is given by $X=(I-PQ)^{-1}P$. We have $X\geqq 0$ if and only if

$$\rho(PQ)<1. \tag{B23}$$

Proof. We first make a transformation on $\tilde{F}(X)$:

$$F_{tr}(X):=\begin{pmatrix} I & (I+XQ)F_0^T\\ 0 & I\end{pmatrix}\tilde{F}(X)\begin{pmatrix} I & 0\\ F_0(I+QX) & I\end{pmatrix} \tag{B24}$$

$$=\begin{pmatrix} \bar{A}^TX+X\bar{A}+\bar{C}_2^T\bar{C}_2+XMX & XB\\ B^TX & D_2^TD_2\end{pmatrix}, \tag{B25}$$

where

$$\bar{A}:=A+BF_0+Q(C_2+D_2F_0)^T(C_2+D_2F_0), \tag{B26}$$

$$\bar{C}_2:=C_2+D_2F_0, \tag{B27}$$

$$M:=(A+BF_0)Q+Q(A^T+F_0^TB^T)+EE^T+Q\bar{C}_2^T\bar{C}_2Q, \tag{B28}$$

and F_0 as defined in (A3). We also transform the second matrix appearing in the rank conditions:

$$\begin{aligned}W(X,s)&:=\begin{pmatrix} I & 0 & -QF_0^T\\ 0 & I & (I+XQ)F_0^T\\ 0 & 0 & I\end{pmatrix}\begin{pmatrix}\tilde{L}(X,s)\\ \tilde{F}(X)\end{pmatrix}\begin{pmatrix} I & 0\\ F_0(I+QX) & I\end{pmatrix}\\ &=\begin{pmatrix} sI-\bar{A}-MX & -B\\ \bar{A}^TX+X\bar{A}+\bar{C}_2^T\bar{C}_2+XMX & XB\\ B^TX & D_2^TD_2\end{pmatrix}.\end{aligned}$$

We have the following equality:

$$\text{(B29)} \qquad \operatorname{normrank} \tilde{G} = \operatorname{normrank} \begin{pmatrix} sI - A_Q & -B_Q \\ C_2 & D_2 \end{pmatrix} - n$$

$$\text{(B30)} \qquad = \operatorname{normrank} \begin{pmatrix} I & QC_2^T \\ 0 & I \end{pmatrix} \begin{pmatrix} sI - A_Q & -B_Q \\ C_2 & D_2 \end{pmatrix} - n$$

$$\text{(B31)} \qquad = \operatorname{normrank} \begin{pmatrix} sI - A & -B \\ C_2 & D_2 \end{pmatrix} - n = \operatorname{normrank} G.$$

Therefore the conditions that $X \geqq 0$ has to satisfy can be reformulated as:

(i) $F_{tr}(X) \geqq 0$,

(ii) rank $F_{tr}(X) = \operatorname{normrank} G$,

(iii) rank $W(X, s) = \operatorname{normrank} G + n \quad \forall s \in \mathscr{C}^0 \cup \mathscr{C}^+$.

Moreover, we note that $T(A, B, C_2, D_2) = T(\bar{A}, B, \bar{C}_2, D_2)$. This can be shown by using the fact that the new system is obtained by a state feedback and an output injection (note that $B = B + Q(C_2 + D_2F_0)^T D_2$) and it is well known that the strongly controllable subspace is invariant under feedback and output injection. This can easily be shown using the algorithm (4.10). We now choose the bases from Appendix A. By Lemma A4(i) we know that if X exists then it will have the form

$$\text{(B32)} \qquad X = \begin{pmatrix} X_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

for some positive semidefinite matrix X_1. Note that there is small difference since M is not necessarily positive semidefinite, but it can be easily seen from the proof in [18] that this difference is not important. We use this decomposition for X and the corresponding decompositions for P and Q:

$$\text{(B33)} \qquad P = \begin{pmatrix} P_1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \qquad Q = \begin{pmatrix} Q_{11} & Q_{12} & Q_{13} \\ Q_{21} & Q_{22} & Q_{23} \\ Q_{31} & Q_{32} & Q_{33} \end{pmatrix}.$$

Together with the decompositions for the other matrices as given in (A4)-(A5) we can decompose $F_{tr}(X)$ correspondingly:

$$\begin{pmatrix} X_1\tilde{A}_{11} + \tilde{A}_{11}^T X_1 + C_{21}^T C_{21} + X_1 M_{11} X_1 & 0 & X_1 \tilde{A}_{13} + C_{21}^T C_{23} & X_1 B_{11} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ \tilde{A}_{13}^T X_1 + C_{23}^T C_{21} & 0 & C_{23}^T C_{23} & 0 & 0 \\ B_{11}^T X_1 & 0 & 0 & \hat{D}_2^T \hat{D}_2 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where

$$\text{(B34)} \qquad \tilde{A}_{11} := A_{11} + Q_{11} C_{21}^T C_{21} + Q_{13} C_{23}^T C_{21},$$

$$\text{(B35)} \qquad \tilde{A}_{13} := A_{13} + Q_{11} C_{21}^T C_{23} + Q_{13} C_{23}^T C_{23},$$

$$\text{(B36)} \qquad M_{11} := A_{11} Q_{11} + A_{13} Q_{13}^T + Q_{11} A_{11}^T + Q_{13} A_{13}^T + E_1 E_1^T$$
$$+ Q_{11} C_{21}^T (C_{21} Q_{11} + C_{23} Q_{13}) + Q_{13} C_{23}^T (C_{21} Q_{11} + C_{23} Q_{13}).$$

The rank condition $\operatorname{rank} F_{tr}(X) = \operatorname{normrank} G$ is, according to Lemma A3(iii), equivalent with the condition that the rank of the above matrix is equal to the rank of the submatrix

$$\text{(B37)} \qquad \begin{pmatrix} C_{23}^T C_{23} & 0 \\ 0 & \hat{D}_2^T \hat{D}_2 \end{pmatrix}.$$

Therefore the Schur complement with respect to this submatrix should be zero. This implies that if we define

$$\begin{aligned} \tilde{R}(X_1) := {} & X_1 \tilde{A}_{11} + \tilde{A}_{11}^T X_1 + C_{21}^T C_{21} + X_1 (M_{11} - B_{11} (\hat{D}_2^T \hat{D}_2)^{-1} B_{11}^T) X_1 \\ & - (X_1 \tilde{A}_{13} + C_{21}^T C_{23})(C_{23}^T C_{23})^{-1} (\tilde{A}_{13}^T X_1 + C_{23}^T C_{21}), \end{aligned}$$

then X_1 should satisfy $\tilde{R}(X_1) = 0$. Moreover, if we decompose $W(X, s)$ correspondingly, then we can show by using elementary row and column operations that for any matrix X in the form (B32), where X_1 satisfies $\tilde{R}(X_1) = 0$, that for all $s \in \mathscr{C}$, $W(X, s)$ has the same rank as the following matrix:

$$\text{(B38)} \qquad \begin{pmatrix} sI - \tilde{Z}(X_1) & 0 & 0 & 0 & 0 \\ * & sI - A_{22} & -A_{23} & 0 & -B_{22} \\ * & -A_{32} & sI - A_{33} & 0 & -B_{32} \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

where

$$\text{(B39)} \quad \tilde{Z}(X_1) := \tilde{A}_{11} + M_{11} X_1 - B_{11} (\hat{D}_2^T \hat{D}_2)^{-1} B_{11}^T X_1 - \tilde{A}_{13} (C_{23}^T C_{23})^{-1} (\tilde{A}_{13}^T X_1 + C_{23}^T C_{21}).$$

The matrix

$$\text{(B40)} \qquad \begin{pmatrix} sI - A_{22} & -A_{23} & -B_{22} \\ -A_{32} & sI - A_{33} & -B_{32} \\ 0 & I & 0 \end{pmatrix}$$

has full row rank for all $s \in \mathscr{C}$ by Lemma A3(ii) and Lemma A1. Hence the rank of the matrix (B38) is $n + \operatorname{normrank} G$ for all $s \in \mathscr{C}^+ \cup \mathscr{C}^0$ if and only if the matrix $\tilde{Z}(X_1)$ is asymptotically stable. Using this we can now reformulate the conditions that $X_1 \geqq 0$ must satisfy:

(i) $\tilde{R}(X_1) = 0$,

(ii) $\tilde{Z}(X_1)$ is asymptotically stable.

That is, X_1 should be the positive semidefinite stabilizing solution of the algebraic Riccati equation $\tilde{R}(X_1) = 0$. Denote the Hamiltonian corresponding to this ARE by H_{new}. We know that P_1 is the stabilizing solution of the algebraic Riccati equation $R(P_1) = 0$ as given by (A9). Denote the Hamiltonian corresponding to this algebraic Riccati equation by H_{old}. Then it can be checked that

$$\text{(B41)} \qquad H_{\text{old}} = \begin{pmatrix} I & Q_{11} \\ 0 & I \end{pmatrix} H_{\text{new}} \begin{pmatrix} I & -Q_{11} \\ 0 & I \end{pmatrix}.$$

Since P_1 is the stabilizing solution of the Riccati equation corresponding to the Hamiltonian H_{old} we know that the modal subspace of H_{old} corresponding to the open left halfplane is given by

$$\text{(B42)} \qquad \mathscr{X}_g(H_{\text{old}}) = \operatorname{Im}\begin{pmatrix} I \\ P_1 \end{pmatrix}.$$

Combining (B41) and (B42), we find

$$\text{(B43)} \qquad \mathscr{X}_g(H_{\text{new}}) = \operatorname{Im}\begin{pmatrix} I & -Q_{11} \\ 0 & I \end{pmatrix}\begin{pmatrix} I \\ P_1 \end{pmatrix} = \operatorname{Im}\begin{pmatrix} I - Q_{11}P_1 \\ P_1 \end{pmatrix}.$$

Therefore we know that there exists a stabilizing solution to the algebraic Riccati equation $\tilde{R}(X_1) = 0$ if and only if $I - Q_{11}P_1$ is invertible and in that case the solution is given by $X_1 = P_1(I - Q_{11}P_1)^{-1}$. This implies that $X = P(I - QP)^{-1} = (I - PQ)^{-1}P$. The requirement $X \geqq 0$ is satisfied if and only if $\rho(PQ) < 1$, which can be checked straightforwardly. This completes the proof. □

Acknowledgment. I thank Harry Trentelman for reading the preliminary draft of this paper and for suggesting many improvements. Moreover, I thank him for his help on the section about the construction of the required compensator.

REFERENCES

[1] J. C. DOYLE, *Lecture notes in advances in multivariable control*, ONR/Honeywell Workshop, Minneapolis, MN, 1984.

[2] J. C. DOYLE, K. GLOVER, P. P. KHARGONEKAR, AND B. A. FRANCIS, *State space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831-847.

[3] B. A. FRANCIS, *The optimal linear-quadratic time invariant regulator with cheap control*, IEEE Trans. Automat. Control, 24 (1979), pp. 616-621.

[4] ———, *A Course in H_∞ Control Theory*, Lecture Notes in Control and Information Sciences, Vol. 88, Springer-Verlag, Berlin, 1987.

[5] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds*, Internat. J. Control, 39 (1984), pp. 1115-1193.

[6] K. GLOVER AND J. C. DOYLE, *State-space formulaes for all stabilizing controllers that satisfy an H_∞ norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167-172.

[7] M. L. J. HAUTUS, *Strong detectability and observers*, Linear Algebra Appl., 50 (1983), pp. 353-368.

[8] P. P. KHARGONEKAR, I. R. PETERSEN, AND M. A. ROTEA, *H_∞ optimal control with state feedback*, IEEE Trans. Automat. Control, 33 (1988), pp. 786-788.

[9] H. KWAKERNAAK, *A polynomial approach to minimax frequency domain optimization of multivariable feedback systems*, Internal. J. Control, 41 (1986), pp. 117-156.

[10] D. J. N. LIMEBEER AND Y. S. HUNG, *An analysis of the pole-zero cancellations in H_∞ optimal control problems of the first kind*, SIAM J. Control Optim., 25 (1987), pp. 1457-1493.

[11] D. J. N. LIMEBEER AND G. D. HALIKIAS, *A controller degree bound for H_∞ control problems of the second kind*, SIAM J. Control Optim., 26 (1988), pp. 646-677.

[12] I. R. PETERSEN AND C. V. HOLLOT, *A Riccati equation approach to the stabilization of uncertain linear system*, Automatica, 22 (1986), pp. 397-411.

[13] I. R. PETERSEN, *Some new results on algebraic Riccati equations arising in linear quadratic differential games and the stabilization of uncertain systems*, Systems Control Lett., 10 (1988), pp. 341-348.

[14] R. M. REDHEFFER, *On a certain linear fractional transformation*, J. Math. Phys., 39 (1960), pp. 269-286.

[15] M. G. SAFONOV, *Imaginary axis zeros in multivariable H_∞ optimal control*, Proc. NATO Workshop on Modelling, Robustness and Sensitivity Reduction in Control Systems, Groningen, 1986.

[16] J. M. SCHUMACHER, *On the structure of strongly controllable systems*, Internat. J. Control, 38 (1983), pp. 525-545.

[17] ———, *A geometric approach to the singular filtering problem*, IEEE Trans. Automat. Control, 30 (1985), pp. 1075-1082.

[18] A. A. STOORVOGEL AND H. L. TRENTELMAN, *The quadratic matrix inequality in singular H_∞ control with state feedback*, SIAM J. Control Optim., 28 (1990), pp. 1190-1208.

[19] A. A. STOORVOGEL, *H_∞ control with state feedback*, in Proc. MTNS-89, Amsterdam, The Netherlands, 1989.

[20] G. TADMOR, *H_∞ in the time domain: the standard four blocks problem*, Mathematics of Control, Signals and Systems, to appear.

[21] H. L. TRENTELMAN, *Almost invariant subspaces and high gain feedback*, CWI Tracts, Vol. 29, Amsterdam, 1986.

[22] G. ZAMES, *Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, 26 (1981), pp. 301-320.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 1, pp. 185-196, January 1991

ASYMPTOTICALLY STABILIZING FEEDBACK CONTROLS AND THE NONLINEAR REGULATOR PROBLEM*

HENRY HERMES†

Abstract. Continuous asymptotically stabilizing feedback controls are constructed for two-dimensional, and certain three-dimensional, small time locally controllable affine systems. The Lie products which determine controllability induce a dilation; it suffices to work with the approximating homogeneous system associated with this dilation. A cost functional is then constructed which is such that the associated Hamilton-Jacobi-Bellman equation is homogeneous, forcing the solution (which is a Lyapunov function for the optimally controlled system) to be homogeneous and thereby determining its basic form. The process may be viewed as a generalization of the linear regulator construction.

Key words. stabilization, controllability, nonlinear control, nonlinear regulator, Hölder continuous feedback

AMS(MOS) subject classification. 93D15

Introduction. Our goal is to develop a method to establish the existence, and exhibit the construction, of continuous asymptotically stabilizing feedback controls for n-dimensional, real analytic, affine control systems of the form

$$\dot{x} = X(x) + uY(x), \qquad X(0) = 0. \tag{1}$$

Our methods use high-order homogeneous approximations and the nonlinear analogue of the classical linear regulator problem.

Brockett [1] gave the following three necessary conditions for the existence of a C^1 feedback control $x \to u(x)$ which makes the solution $x(t) = 0$ of (1) locally asymptotically stable.

(B1) The linearized system should have no uncontrollable modes associated with eigenvalues whose real part is positive.

(B2) For each y in a nbd of zero there exists an open-loop control $t \to u_y(t)$ such that the corresponding solution of (1) initiating from y tends to zero as $t \to \infty$.

(B3) The map $\gamma: \mathbb{R}^n \times \mathbb{R}^1 \to \mathbb{R}^n$ defined by $\gamma(x, u) = X(x) + uY(x)$ must be onto a nbd of zero.

Prior to Brockett's paper, Sontag and Sussmann [2] showed that certain nonlinear systems could have rest solutions made asymptotically stable by continuous, but not C^1, feedback control.

Kawski [3] showed that condition (B1) is not necessary for the existence of a C^0 asymptotically stabilizing feedback control for a two-dimensional system of the form (1). Indeed, he showed that the zero solution of the system $\dot{x}_1 = u$, $\dot{x}_2 = x_2 - x_1^3$ could be made locally asymptotically stable via feedback of the form $u(x) = (4E/3)x_2^{1/3} - x_1 - K(x_1^3 - x_2)$.

System (1) is said to be small time locally controllable (at zero), denoted STLC, if for any $t_1 > 0$ the set of points attainable at time t_1 by solutions initiating from zero, corresponding to all admissible open-loop controls $t \to u(t)$, contains a full nbd of zero. By reversing time, clearly if (1) is STLC, then the system $\dot{x} = -X(x) + u(Y)(x)$

* Received by the editors August 7, 1989; accepted for publication (in revised form) January 31, 1990.

† Department of Mathematics, Box 426, University of Colorado, Boulder, Colorado 80309-0426. This research was supported by National Science Foundation grant DMS-8721917.

is such that all points in some nbd of zero can be steered to zero in arbitrarily small time. Sufficient conditions for STLC (see [4]-[6]), when satisfied for system (1), also imply STLC of the reversed time system. Indeed, in the real analytic case, Sussmann has shown STLC of system (1) and the reversed time system are equivalent. Hence STLC of system (1) is sufficient for property (B2). We shall assume STLC throughout, although it is by no means necessary for (B2).

An outline of a proof of the necessity of condition (B3) via fixed point theory is given in [1]; a proof in the continuous category, based on index theory, is given in [7]. It follows that (B3) is therefore a necessary condition for the existence of a continuous, i.e., C^0, asymptotically stabilizing feedback control. For dimension $n=2$, STLC implies (B3).

Kawski's main result in [3] is the following theorem.

THEOREM 1 (Kawski [3]). *If the two-dimensional, real analytic system* (1) *is* STLC, *then there exists a continuous feedback control which makes the zero solution locally asymptotically stable.*

His method is via the construction of a Lyapunov function V having trajectory derivative $\dot{V}$ negative semidefinite along solutions of the controlled system and such that (essentially) LaSalle's theorem applies on the set where $\dot{V}=0$. The methods are limited to dimension $n=2$. Our first task here is to illustrate the ideas involved in the nonlinear regulator method by using these to prove Theorem 1. The resulting construction of a feedback control and Lyapunov function V (see Lemma 2) is such that when applied to the example $\dot{x}_1=u$, $\dot{x}_2=x_2-x_1^3$ gives the trajectory derivative $\dot{V}(x)$ negative definite. The proof of Lemma 2 contains the main ideas of the nonlinear regulator approach and of this paper.

Many examples and results for two-dimensional systems, some related to the above and others for systems which are not STLC, can be found in [8]. The use of center manifold theory for establishing smooth asymptotically stabilizing feedback controls for two-dimensional systems can be found in [9].

In dimension $n \geqq 3$ much less is known. The system $\dot{x}_1=u$, $\dot{x}_2=x_1$, $\dot{x}_3=-x_1^3$ is STLC but clearly does not satisfy condition (B3). Specifically, the point $(0,0,\varepsilon)$, $\varepsilon \neq 0$, is not in the range of the map γ, hence this system does not admit a continuous asymptotically stabilizing feedback control. In § 2 we utilize the nonlinear regulator method to show the (surprising) result that the seemingly more difficult to control system $\dot{x}_1=u$, $\dot{x}_2=x_1$, $\dot{x}_3=x_3-x_1^3$ does admit a continuous, asymptotically stabilizing, feedback control.

1. Homogeneity and the nonlinear regulator.

The constructions that make the linear regular problem tractable, e.g., the Riccati equation, etc., result from homogeneity properties of the linear system and quadratic cost functional involved. We will elaborate on this comment shortly, but first introduce the more general notions of homogeneity. A *dilation* $\delta_\varepsilon^r : \mathbb{R}^n \to \mathbb{R}^n$ will be a map of the form $\delta_\varepsilon^r x = (\varepsilon^{r_1}x_1, \cdots, \varepsilon^{r_n}x_n)$ where $x=(x_1,\cdots,x_n)$ are given local coordinates, $1=r_1 \leqq \cdots \leqq r_n$ are integers, and $\varepsilon > 0$. If all $r_i=1$ we write δ_ε^1 and refer to this as the *standard dilation.* A function $h:\mathbb{R}^n \to \mathbb{R}^1$ is homogeneous of degree s with respect to δ_ε^r if $h(\delta_\varepsilon^r x)=\varepsilon^s h(x)$. Functions homogeneous of degree s will be denoted by H_s with H_0 denoting constant functions and $H_s=\{0\}$ if $s<0$. If $h \in H_s$, $g \in H_q$, it follows that $\partial h/\partial x_j \in H_{s-r_j}$ while the product $hg \in H_{s+q}$. A vector field X on $\mathbb{R}^n$ is homogeneous of degree s with respect to δ_ε^r if $Xh \in H_{q-s}$ whenever $h \in H_q$. Thus if in local coordinates $X(x)=\sum_{j=1}^n a_j(x)\partial/\partial x_j$, X is homogeneous of degree s means $a_j \in H_{r_j-s}$. This has (sadly) become standard usage but the reader should be warned that a vector field $X(x)=\sum_{j=1}^n a_j(x)\partial/\partial x_j$, which is homogeneous of degree m with respect to δ_ε^1 in the classical sense (i.e., $a_i(\varepsilon x) = \varepsilon^m a_i(x)$), now becomes homogeneous of degree $(1-m)$ in the present sense.

1.1. The linear regulator. The well-known "pole placement" theorems show that a controllable linear system on $\mathbb{R}^n$

$$\dot{x} = Ax + Bu, \qquad u \in \mathbb{R}^k \tag{2}$$

can have its zero solution made globally asymptotically stable via linear feedback $u^*(x) = Kx$. This was first established by Kalman [10] in his study of the linear regulator problem. Specifically, consider the controllable system (2) with cost functional (to be minimized)

$$C(u) = \int_0^\infty (u'Ru + x'Wx)\,dt, \qquad R, W \text{ positive, definite, symmetric.}$$

Then for any $x(0) = x^0 \in \mathbb{R}^n$ there is a solution $\varphi(t, x^0)$ with $\varphi(t, x^0) \to 0$ as $t \to \infty$ and the "cost, or value, from x^0" denoted $V(x^0)$ is defined and continuous on $\mathbb{R}^n$ (see [11, p. 198]). Following Kalman [10] for $p \in \mathbb{R}^n$ form

$$H(x, p, u) = p'Ax + p'Bu + u'Ru + x'Wx. \tag{3}$$

Minimize $H(x, p, u)$ with respect to u. Since there are no bounds on the control values, this requires $H_u(x, p, u) = 2Ru + B'p = 0$ or

$$u^*(x, p) = (\tfrac{1}{2})R^{-1}B'p.$$

Next, let $H^*(x, p) = H(x, p, u^*(x, p))$ and the value function V must satisfy the Hamilton-Jacobi-Bellman (HJB) equation

$$H^*(x, V_x(x)) = 0, \quad V(0) = 0, \quad V(x) > 0 \quad \text{if } x \neq 0. \tag{4}$$

Now the homogeneity of the problem comes into play. Consider $(x_1, \cdots, x_n, V, p_1, \cdots, p_n)$ as local coordinates for the first jet bundle $J^1(\mathbb{R}^n, \mathbb{R}^1)$. With any dilation $\delta_\varepsilon^r x = (\varepsilon^{r_1}x_1, \cdots, \varepsilon^{r_n}x_n)$ and integer $m > r_n$ associate the dual dilation $\delta_\varepsilon^{r*}p = (\varepsilon^{m-r_1}p_1, \cdots, \varepsilon^{m-r_n}p_n)$ and dilation γ_ε^r on $J^1(\mathbb{R}^n, \mathbb{R}^1)$ defined by $\gamma_\varepsilon^r(x, V, p) = (\delta_\varepsilon^r x, \varepsilon V, \delta_\varepsilon^{r*}p)$. If $H^*(x, p)$ is homogeneous of degree m with respect to γ_ε^r, i.e., $H^*(\delta_\varepsilon^r x, \delta_\varepsilon^{r*}p) = \varepsilon^m H^*(x, p)$, we expect a solution V of the HJB equation which is homogeneous of degree m.

In the linear problem, letting δ_ε^r be the standard dilation δ_ε^1 and $m = 2$ (so δ_ε^{r*} is also the standard dilation) it follows that $H^*(x, p)$ is homogeneous of degree two and we should seek a solution of (4) of the form

$$V(x) = x'Ex, \tag{5}$$

where, without loss of generality, E may be assumed to be symmetric. Substitution of this V into (4) yields $2x'EAx - x'EBR^{-1}B'Ex + x'Wx = 0$ or that E should satisfy the "stationary Riccati equation"

$$A'E + EA - EBR^{-1}B'E = -W. \tag{6}$$

Given positive-definite matrices R, W, (6) determines a unique positive-definite symmetric matrix E. The solution of (4) is $V(x) = x'Ex$ and the optimal feedback control (which provides global asymptotic stability of (2)) is

$$u^*(x) = -\tfrac{1}{2}R^{-1}B'V_x(x) = -R^{-1}B'Ex. \tag{7}$$

The well-known results above were repeated since they form an example and a guide for the nonlinear regulator problem which follows.

1.2. The nonlinear regulator. For the linear regulator, a positive-definite homogeneous (of degree two) form with respect to the standard dilation δ_ε^1 was known to have the representation $V(x)=x'Ex$ with E a positive-definite matrix. The next lemma gives elementary conditions which tell when forms $V(x)$, which are homogeneous of even degree with respect to an arbitrary dilation, are positive definite. We denote by $\alpha=(\alpha_1,\cdots,\alpha_n)$ a multi-index with all α_i nonnegative rationals and let $|x|^\alpha=|x_1|^{\alpha_1}\cdots|x_n|^{\alpha_n}$. Then $|x|^\alpha$ is homogeneous of degree m with respect to $\delta_\varepsilon^r x=(\varepsilon^{r_1}x_1,\cdots,\varepsilon^{r_n}x_n)$ if and only if $\sum_{i=1}^n \alpha_i r_i=m$.

LEMMA 1. *Let $\delta_\varepsilon^r x=(\varepsilon^{r_1}x_1,\cdots,\varepsilon^{r_n}x_n)$ be a given dilation and $V(x)=\sum C_\alpha|x|^\alpha$ with the sum taken over a finite set of multi-indices $\alpha=(\alpha_1,\cdots,\alpha_n)$ which satisfy $\sum_{i=1}^n \alpha_i r_i=m$, m an even integer. Thus V is homogeneous of degree m with respect to δ_ε^r. Call a multi-index α mixed if more than one α_i is nonzero and assume the nonmixed multi-indices $\alpha^1=(m/r_1,0,\cdots,0),\cdots,\alpha^n=(0,\cdots,m/r_n)$ all appear in the sum defining V and have corresponding coefficients $C_{\alpha_i}=E_i>0$, $i=1,\cdots,n$. If*

$$\sum|C_\alpha|\left(\frac{1}{E_1}\right)^{\alpha_1 r_1/m}\cdots\left(\frac{1}{E_n}\right)^{\alpha_n r_n/m}<1 \tag{8}$$

with the sum taken over all mixed α in the sum defining V, then V is positive definite on $\mathbb{R}^n$.

Proof. Let E denote $(E_1,\cdots,E_n)$ and define

$$\Gamma_E=\{x\in\mathbb{R}^n\colon E_1|x_1|^{m/r_1}+E_2|x_2|^{m/r_2}+\cdots+E_n|x_n|^{m/r_n}=1\}.$$

Then Γ_E is a closed hypersurface which "encloses" the origin. For $x\in\Gamma_E$, $|x_1|\leqq(1/E_1)^{r_1/m},\cdots,|x_n|\leqq(1/E_n)^{r_n/m}$ and hence $|x_i|\to 0$ as E_i increases. We write $V(x)=\sum_{i=1}^n E_i|x_i|^{m/r_i}+\sum C_\alpha|x|^\alpha$ with the latter sum taken over mixed α occurring in the original sum defining V. Thus if

$$\sum|C_\alpha||x|^\alpha\leqq\sum|C_\alpha|\left(\left(\frac{1}{E_1}\right)^{\alpha_1 r_1/m}\cdots\left(\frac{1}{E_n}\right)^{\alpha_n r_n/m}\right)<1$$

we see $V(x)>0$ on Γ_E. Since $V\in H_m$, i.e., $V(\delta_\varepsilon^r x)=\varepsilon^m V(x)$, this implies V is positive definite on $\mathbb{R}^n$. □

Example 1.1. Let $n=2$, $\delta_\varepsilon^r x=(\varepsilon x_1,\varepsilon^3 x_2)$, $m=4$, and $V(x)=16x_1^4+Cx_1x_2+x_2^{4/3}$. Then $V\in H_4$, $E_1=16$, $E_2=1$, and the only mixed $\alpha=(1,1)$. Thus if $|C|(1/16)^{1/4}(1/1)^{3/4}<1$, or $|C|<2$, V is positive definite.

Example 1.2 (the lemma is not sharp). Take $n=2$, $\delta_\varepsilon^r x=(\varepsilon x_1,\varepsilon x_2)$, $m=2$, and $V(x)=x_1^2+Cx_1x_2+x_2^2$. Then $E_1=E_2=1$; the only mixed $\alpha=(1,1)$, and the lemma gives V positive definite if $|C|(1/1)^{1/2}(1/1)^{1/2}<1$ or $|C|<1$. Clearly, $|C|<2$ suffices.

Proof of Theorem 1. Since the two-dimensional system (1) was assumed STLC, $Y(0)\neq 0$ and we can choose local coordinates so that $Y(0)=\partial/\partial x_1$. If, in these coordinates $X(x)=a_1(x)\partial/\partial x_1+a_2(x)\partial/\partial x_2$, choose feedback $u(x)=-a_1(x)+\mu(x)$ so (renaming μ as u) with no loss of generality we may assume system (1) has the form

$$\dot{x}_1=u,\quad \dot{x}_2=a_2(x),\quad a_2(0)=0;$$

i.e., $X(x)=a_2(x)\partial/\partial x_2$, $Y=\partial/\partial x_1$. From [12], STLC at zero implies there is an integer k such that the Lie product $(ad^kY,X)(0)=(0,\partial^k a_2(0)/\partial x_1^k)\neq 0$ and that the smallest such k must be odd. This means that $a_2(x)=x_1^k q(x)+p(x)$ where $q(0)=C_0\neq 0$ and $\partial^j p(0)/\partial x_1^j=0, j=0,\cdots,k$. Thus a canonical form for a two-dimensional STLC system of the form (1) may be taken as

$$\dot{x}_1=u,\qquad \dot{x}_2=x_1^k q(x)+p(x) \tag{9}$$

with k odd and p, q, as above.

With system (9) associate the dilation $\delta_\varepsilon^r x(\varepsilon x_1, \varepsilon^k x_2)$ and expand q, p in homogeneous polynomials relative to this dilation, i.e., from the form of p, q

$$q(x) = C_0 + \sum_{j=1}^{\infty} q_j(x), \qquad q_j \in H_j,$$

$$p(x) = C_1 x_2 + \sum_{j=k+1}^{\infty} p_j(x), \qquad p_j \in H_j.$$

We write (9) as

$$\dot{x}_1 = u, \qquad \dot{x}_2 = C_0 x_1^k + C_1 x_2 + R(x), \tag{10}$$

where

$$R(x) = \sum_{j=1}^{\infty} x_1^k q_j(x) + \sum_{j=k+1}^{\infty} p_j(x)$$

and thus the vector field $R(x)\partial/\partial x_2$ is a sum of vector fields homogeneous of degree $\leqq -1$ with respect to δ_ε^r. We shall call

$$\dot{x}_1 = u, \qquad \dot{x}_2 = C_0 x_1^k + C_1 x_2, \qquad C_0 \neq 0, \tag{11}$$

the approximating system.

The proof will proceed by showing the following lemma.

LEMMA 2. *There exists a continuous feedback control $u^* \in H_1$ for the approximating system* (11) *such that*

(a) *Solutions to* (11) *with $u = u^*(x)$ are unique.*

(b) *The zero solution of* (11) *with $u = u^*(x)$ is globally asymptotically stable.*

(c) *There is a positive-definite Lyapunov function $V \in H_{k+1}$ such that its trajectory derivative $\dot{V}(x)$ along solutions of* (11), *with $u = u^*(x)$, is negative definite.*

The vector field $u^*(x)\partial/\partial x_1 + (C_0 x_1^k + C_1 x_2)\partial/\partial x_2$ is then homogeneous of degree zero with respect to δ_ε^r while the "remainder" vector field $R(x)\partial/\partial x_2$ is a sum of vector fields homogeneous of degrees less than or equal to -1. The conclusion of the proof of Theorem 1 then follows from Theorem 2.

THEOREM 2 [13, Thm. 1]. *Let W be a continuous vector field on $\mathbb{R}^n$ with $W(0) = 0$ which is homogeneous of degree m with respect to a dilation $\delta_\varepsilon^r x = (\varepsilon^{r_1} x_1, \cdots, \varepsilon^{r_n} x_n)$ and such that solutions to initial value problems for $\dot{x} = W(x)$ are unique. Let Z be a continuous sum of vector fields homogeneous of degree less than or equal to $(m-1)$. If the zero solution of $\dot{x} = W(x)$ is locally asymptotically stable, then the same is true for the zero solution of $\dot{x} = W(x) + Z(x)$.*

Proof of Lemma 2. Throughout, the dilation is $\delta_\varepsilon^r x = (\varepsilon x_1, \varepsilon^k x_2)$, k odd and determined by the Lie products which showed STLC. With system (11) consider the optimization problem of reaching the origin while minimizing the cost functional

$$C(u) = \int_0^\infty [e u^{k+1}(s) + h_{k+1}^+(x(s))\, ds], \qquad e > 0$$

with $h_{k+1}^+ \in H_{k+1}$ positive definite and free to be chosen later. Following the linear regulator approach, define

$$H(x, p, u) = p_1 u + C_0 p_2 x_1^k + C_1 x_2 + e u^{k+1} + h_{k+1}^+(x).$$

Since the values of u are unconstrained, minimizing H with respect to u yields

$$u^*(x, p) = -\left(\frac{p_1}{e(k+1)}\right)^{1/k}. \tag{12}$$

Define

$$H^*(x,p)=H(x,p,u^*(x,p))=-\gamma p_1^{(k+1)/k}+C_0p_2x_1^k+C_1p_2x_2+h_{k+1}^+(x), \tag{13}$$

where

$$\gamma=\left(\frac{k}{k+1}\right)\left(\frac{1}{e(k+1)}\right)^{1/k}>0. \tag{14}$$

Note that if we define the dual dilation to δ_ε^r as $\delta_\varepsilon^{r*}p=(\varepsilon^{k+1-r_1}p_1,\varepsilon^{k+1-r_2}p_2)=(\varepsilon^kp_1,\varepsilon p_2)$, then with $\gamma_\varepsilon^r(x,V,\dot{p})=(\delta_\varepsilon^r x,\varepsilon V,\delta_\varepsilon^{r*}p)$ as a dilation on $J^1(\mathbb{R}^2,\mathbb{R}^1)$ we have $H^*(x,p)$ homogeneous of degree $(k+1)$ with respect to γ_ε^r.

The HJB equation is

$$0=H^*(x,V_x(x))=-\gamma V_{x_1}^{(k+1)/k}+C_0x_1^kV_{x_2}+C_1x_2V_{x_2}+h_{k+1}^+(x). \tag{15}$$

The interpretation of this equation is that a solution V is such that $V(x)$ is the optimal cost to reach the origin starting from x. Since the integrand in the cost functional is positive definite, we seek a positive-definite solution V of (15). Since (11) is STLC we know (e.g., [14]) that (15) has a continuous positive-definite viscosity solution. The homogeneity of $H^*(x,p)$ with respect to γ_ε^r implies we should seek a solution $V\in H_{k+1}$ with respect to δ_ε^r. For this two-dimensional case the Brockett necessary condition (B3) is automatically satisfied and we shall show that (11) has a C^1, positive-definite solution $V\in H_{k+1}$. Specifically, since $h_{k+1}^+\in H_{k+1}$ has been left free, we actually show that there exists a C^1, positive-definite $V\in H_{k+1}$ such that

$$-\gamma V_{x_1}^{(k+1)/k}(x)+C_0x_1^kV_{x_2}(x)+C_1x_2V_{x_2}(x) \tag{16}$$

is negative definite. From (12), this will yield

$$u^*(x)=-\left(\frac{V_{x_1}(x)}{e(k+1)}\right)^{1/k} \tag{17}$$

as the desired continuous feedback control.

Since we expect $V\in H_{k+1}$ and $V\in C^1$, choose the form of V as

$$V(x)=\beta_1x_1^{k+1}+\beta_2x_1x_2+2\beta_3x_2^{(k+1)/k},\ \beta_1,\beta_2>0. \tag{18}$$

Note that terms such as $x_1^kx_2^{1/k}$, $x_1^{k-1}x_2^{2/k}$ are in H_{k+1} but were omitted since they would not yield a C^1 function V. From Lemma 1, V is positive definite if

$$|\beta_2|<(\beta_1)^{1/(k+1)}(\beta_3)^{k/(k+1)}. \tag{19}$$

Substituting V, as given in (18), into (16) and calling $g(x)$ the result yields

$$\begin{aligned}g(x)={}&-\gamma[(k+1)\beta_1x_1^k+\beta_2x_2]^{(k+1)/k}+C_0\beta_2x_1^{k+1}\\&+C_0\left(\frac{k+1}{k}\right)\beta_3x_1^kx_2^{1/k}+C_1\beta_1x_1x_2+C_1\left(\frac{k+1}{k}\right)\beta_3x_2^{(k+1)/k}.\end{aligned} \tag{20}$$

Our goal is to choose β_1, β_2, β_3, and $e>0$; hence $\gamma>0$, so that (19) is satisfied and $g\in H_{k+1}$ is negative definite.

Make the local coordinate change (which preserves homogeneity)

$$y_1=x_1,\qquad y_2=2(k+1)\beta_1x_1^k+\beta_2x_2$$

and inverts as

$$x_1=y_1,\qquad x_2=\left(\frac{1}{\beta_2}\right)(y_2-(k+1)\beta_1y_1^k),\quad \beta_2\neq 0.$$

With a slight abuse of notation by again using g for the resulting function, and adding and subtracting y_1^{k+1}, we obtain

$$\begin{aligned} g(y) = & -y_1^{k+1} - \gamma y_2^{(k+1)/k} + (1 + C_0\beta_2 - C_1(k+1)\beta_1)y_1^{k+1} + C_1 y_1 y_2 \\ & + C_0\left(\frac{k+1}{k}\right)\beta_3\left[\left(\frac{1}{\beta_2}\right)(y_2 - (k+1)\beta_1 y_1^k)\right]^{1/k} y_1^k \\ & + C_1\frac{(k+1)}{k}\beta_3\left[\left(\frac{1}{\beta_2}\right)(y_2 - (k+1)\beta_1 y_1^k)\right]^{(k+1)/k}. \end{aligned}$$

Again, $g(\delta_\lambda^r y) = \lambda^{k+1} g(y)$. Define

$$\Gamma = \{y \colon y_1^{k+1} + \gamma y_2^{(k+1)/k} = 1\}.$$

Then Γ is a closed curve which encircles the origin. Since $g \in H_{k+1}$ if we can choose $\beta_1, \beta_2, \beta_3$, and $\gamma > 0$ to make g negative on Γ, it follows that g will be negative definite. But on Γ, $|y_1| \leqq 1$ and $|y_2| \leqq 1/\gamma$; hence $y_2 \to 0$ as $\gamma \to \infty$ or equivalently as $e \to 0$. (Intuitively, $e \to 0$ means we weight the cost of control less.) Thus for $y \in \Gamma$ and $e > 0$ sufficiently small, terms in $g(y)$ having a factor y_2 are insignificant and we omit these. For this case

$$g(y) \cong -1 + f(\beta) y_1^{k+1}, \qquad |y_1| \leqq 1.$$

It suffices to show that we can choose β_1, β_2, β_3 to satisfy (19) and so that $f(\beta) < 1$ or $F(\beta) = f(\beta) - 1 < 0$ where

$$\begin{aligned} F(\beta) = & C_0\beta_2 - C_1(k+1)\beta_1 - C_0\left(\frac{k+1}{k}\right)\beta_3\left[\left(\frac{\beta_1}{\beta_2}\right)(k+1)\right]^{1/k} \\ & \cdot \left[1 - \left(\frac{C_1}{C_0}\right)\frac{(k+1)\beta_1}{\beta_2}\right]. \end{aligned} \tag{21}$$

To see that this can always be done, first choose $\beta_1 = 1$; next choose β_2 so that $\beta_2 C_0 > 0$ (recall $C_0 \neq 0$ in (11)) and $|\beta_2|$ sufficiently large so that $C_1(k+1)/(C_0\beta_2) < 1$. From (21) it is then clear that for $\beta_3 > 0$ and sufficiently large, $F(\beta) < 0$. Finally, by increasing β_3 if necessary (i.e., assure $\beta_3^{3/4} > |\beta_2|$) we can satisfy (19).

With the choices as above, $V(x)$ as given by (18) is positive definite while with $u = u^*(x)$ as given by (17), the trajectory derivative of V along solutions of (11) is $\dot{V}(x) = g(x)$ which is negative definite.

Finally, we show that we have uniqueness of solutions of (11) even though $u^*(x)$ is only continuous. With $u = u^*(x)$ (11) becomes $\dot{x}_1 = -(V_{x_1}(x)/(k+1)e)^{1/k}$, $\dot{x}_2 = C_0 x_1^k + C_1 x_2$ and the only difficulties can occur on the curve $\Lambda = \{x \colon V_{x_1}(x) = 0\}$. But on Λ, the vector field is

$$C_0\left(\left(1 - \frac{(k+1)\beta_1 C_1}{C_0\beta_2}\right)x_1^k\right)\frac{\partial}{\partial x_2}$$

and our choice of β_1, β_2 was such that $(1 - ((k+1)\beta_1 C_1/C_0\beta_2)) \neq 0$ while also $C_0 \neq 0$. Since a normal to Λ is $\eta = (k(k+1)\beta_1 x_1^k, \beta_2)$, it follows that for $x \neq 0$ the vector field defined by the right side of (11) is transverse to Λ. By Theorem 14.1 of [15], this means that solutions of (11) are unique. □

The control u^* constructed in Lemma 2 provides *global* asymptotic stability for the zero solution of the approximating system (11) and, by Theorem 2, local asymptotic stability for the zero solution of the original system having (11) as its approximation.

Example 1.3. For the Kawski example $\dot{x}_1 = u$, $\dot{x}_2 = x_2 - x_1^3$ we have $k = 3$, $C_0 = -1$, $C_1 = 1$. To construct V, first choose $\beta_1 = 1$; next, since $C_0 < 0$, we want $B_2 < 0$ and such that $-4/\beta_2 < 1$. Choose $\beta_2 = -32$ in which case (with β_3 still free), from (21), $F(\beta) = 28 - (7/12)\beta_3$. Thus $\beta_3 > 36$ suffices to make $F(\beta) < 0$ while use of Lemma 2.1, or (19), requires $\beta_3^{3/4} > |\beta_2| = 32$, or $\beta_3 > 101.6$. Thus, for example,

$$V(x) = x_1^2 - 32x_1x_2 + 200x_2^{4/3}$$

is positive definite and for $e > 0$ sufficiently small,

$$u^*(x) = -\left(\frac{1}{2e^{1/3}}\right)(2x_1 - 32x_2)^{1/3}$$

is an asymptotically stabilizing feedback control for which the trajectory derivative $\dot{V}(x)$ is negative definite.

Example 1.4 (linear regulator or $k = 1$). With $k = 1$, $C_0 = C_1 = 1$ in (11) we have the linear system $\dot{x}_1 = u$, $\dot{x}_2 = x_1 + x_2$, or $\dot{x} = Ax + bu$, $A = \left(\begin{smallmatrix} 0 & 0 \\ 1 & 1 \end{smallmatrix}\right)$, $b = \left(\begin{smallmatrix} 1 \\ 0 \end{smallmatrix}\right)$. To construct an asymptotically stabilizing feedback control choose $\beta_1 = 1$; then in (21) we want $\beta_2 > 0$ and such that $2/\beta_2 < 1$ or $\beta_2 > 2$. Choose $\beta_2 = 4$ which gives $F(\beta) = 2 - \beta_3/2$ or $\beta_3 > 4$ makes $F(\beta) < 0$. From (19) we need $2 < \beta_1^{1/2}\beta_3^{1/2}$ or $\beta_3 > 4$. Choose $\beta_3 = 8$, hence $V(x) = x_1^2 + 4x_1x_2 + 8x_2^2$ is positive definite and $u^*(x) = -(1/2e)(2x_1 + 4x_2)$ is an asymptotically stabilizing feedback control for $e > 0$ sufficiently small. Here (referring to the cost functionals for the linear regulator) $R = e$ while $V(x) = x'Ex$ with $E = \left(\begin{smallmatrix} 1 & 2 \\ 2 & 8 \end{smallmatrix}\right)$. We are assured that for $e > 0$ sufficiently small

$$(A'E + EA) - \left(\frac{1}{e}\right)(Ebb'E) = -W, \quad W \text{ positive definite} \tag{22}$$

since the trajectory derivative now is $\dot{V}(x) = g(x) = -x'Wx$. Computing in (22) yields

$$\begin{pmatrix} 4 & 10 \\ 10 & 16 \end{pmatrix} - \left(\frac{1}{e}\right)\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} = -W$$

and for, say, $e = \frac{1}{5}$, we do obtain that $W = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 4 \end{smallmatrix}\right)$ is positive definite.

2. A three-dimensional example. Theorem 1, and the construction of an asymptotically stabilizing feedback control for a two-dimensional system of the form (1) which is STLC, is a consequence of the ability to characterize canonical (nilpotent) approximating systems, i.e., (11), for such systems. For $n \geqq 3$, such canonical approximations have not yet been classified; hence, we deal with the specific following example.

Example 2.1. For $n = 3$ consider the system of the form (1) having

$$X(x) = x_1\frac{\partial}{\partial x_2} + (cx_3 - x_1^3)\frac{\partial}{\partial x_3}, \qquad Y = \frac{\partial}{\partial x_1}. \tag{23}$$

As remarked in the Introduction, if $c = 0$, this system does not satisfy condition (B3) and hence does not admit a continuous asymptotically stabilizing feedback control. However, if $c \neq 0$ (we shall assume $c > 0$ which is the interesting case), condition (B3) is satisfied and we shall construct a continuous asymptotically stabilizing control.

Let $\mathcal{S}^1 = \{(ad^\nu X, Y) : \nu = 0, 1, \cdots\}$, $\mathcal{S}^j$ denote the set of all Lie products if i-tuples of elements of $\mathcal{S}^1$ with $i \leqq j$, and $\mathcal{S}^j(0)$ denote the elements of $\mathcal{S}^j$ evaluated at zero. Then $\dim \operatorname{span} \mathcal{S}^1(0) = 2 = \dim \operatorname{span}\{Y(0), [X, Y](0)\}$; $\dim \operatorname{span} \mathcal{S}^2(0) = 2$ while $(ad^3 Y, X)(0) \in \mathcal{S}^3(0)$ is independent of span $\mathcal{S}^2(0)$; i.e., dim span $\mathcal{S}^3(0) = 3$. The system (23) is, in the terminology of [4] and [5], an odd system since dimension span $\mathcal{S}^j(0)$ increases only for odd integers $j$and is STLC for any value of c. The dilation associated with a three-dimensional system of the form (1) having $\dim \operatorname{span} \mathcal{S}^1(0)=2$, $\dim \operatorname{span} \mathcal{S}^2(0)=2$, $\dim \operatorname{span} \mathcal{S}^3(0)=3$ is $\delta_\varepsilon^r x=(\varepsilon x_1, \varepsilon x_2, \varepsilon^3 x_3)$. Our construction will show, by Theorem 2, local asymptotic stability of the zero solution of any system of the form (1) having system (23), with $c \neq 0$, as its (nilpotent) approximating system relative to δ_ε^r.

With system (23), consider the optimization problem of minimizing $C(u)=\int_0^\infty [eu^4(s)+h_4^+(x(s))]\,ds$, $e>0$, $h_4^+ \in H_4$ and positive definite. Then, with the terminology as in the proof of lemma 2,

$$H(x,p,u)=p_1u+p_2x_1+cp_3x_3-p_3x_1^3+eu^4+h_4^+(x),$$

$$u^*(x,p)=-\left(\frac{p_1}{4e}\right)^{1/3},$$

$$H^*(x,p)=-\gamma p_1^{4/3}+p_2x_1+cp_3x_3-p_3x_1^3+h_4^+(x)$$

with

$$\gamma=\left(\frac{3}{4}\right)\left(\frac{1}{e}\right)^{1/3}>0.$$

The HJB equation is $H^*(x, V_x(x))=0$, with data $V(0)=0$, V positive definite. We expect $V \in H_4$. Thus we try

$$V(x)=E_1x_1^4+\beta_1x_1x_2^3+\beta_2x_1x_3+E_2x_2^4+\beta_3x_2x_3+E_3x_3^{4/3} \tag{24}$$

and require E_1, E_2, $E_3>0$. Lemma 1 shows that V is positive definite if

$$|\beta_1|\left(\frac{1}{E_1}\right)^{1/4}\left(\frac{1}{E_2}\right)^{3/4}+|\beta_2|\left(\frac{1}{E_1}\right)^{1/4}\left(\frac{1}{E_3}\right)^{3/4}+|\beta_3|\left(\frac{1}{E_2}\right)^{1/4}\left(\frac{1}{E_3}\right)^{3/4}<1. \tag{25}$$

The HJB "inequality," i.e., the analogue of (16), which we wish to have negative definite is

$$-\gamma V_{x_1}^{4/3}(x)+x_1V_{x_2}(x)+cx_3V_{x_3}(x)-x_1^3V_{x_3}(x). \tag{26}$$

Substituting from (24) into (26) gives

$$\begin{aligned}g(x)=&-\gamma(4E_1x_1^3+\beta_1x_2^3+\beta_2x_3)^{4/3}+3\beta_1x_1^2x_2^2+4E_2x_1x_2^3+(\beta_3+c\beta_2)x_1x_3\\&+c\beta_3x_2x_3+(\tfrac{4}{3})cE_3x_3^{4/3}-\beta_2x_1^4-\beta_3x_1^3x_2-(\tfrac{4}{3})E_3x_1^3x_3^{1/3}.\end{aligned}$$

Change coordinates to

$$y_1=x_1,\quad y_2=x_2,\quad y_3=4E_1x_1^3+\beta_1x_2^3+\beta_2x_3,$$

which inverts as

$$x_1=y_1,\quad x_2=y_2,\quad x_3=\left(\frac{1}{\beta_2}\right)(y_3-4E_1y_1^3-\beta_1y_2^3),$$

where we assume $\beta_2 \neq 0$. Again, abusing notation by retaining the symbol g after the variable change, and from symmetry considerations choosing $\beta_3=\beta_2\neq 0$, we obtain

$$\begin{aligned}g(y)=&-\gamma y_3^{4/3}-(4cE_1+\beta_2+4E_1)y_1^4-c\beta_1y_2^4+(4E_2-\beta_1-c\beta_1)y_1y_2^3\\&+3\beta_1y_1^2y_2^2-(\beta_2+4c)y_1^3y_2+2(c+1)y_1y_3+cy_2y_3\\&+\left(\frac{4}{3}\right)E_3\left(\frac{1}{\beta_2}\right)^{1/3}(y_3-4E_1y_1^3-\beta_1y_2^3)^{1/3}\left[\left(\frac{c}{\beta_2}\right)(y_3-4E_1y_1^3-\beta_1y_2^3)-y_1^3\right].\end{aligned}$$

Our goal is to show that β_1, β_2, E_1, E_2, E_3 can be chosen so that V is positive definite and the homogeneous function $g \in H_4$ is negative definite. Let

(i) $E_2 = 16$, $\beta_1 > 0$, $4cE_1 + \beta_2 + 4E_1 > 0$ and define

$$\Gamma = \{y\colon (4cE_1 + \beta_2 + 4E_1)y_1^4 + c\beta_1 y_2^4 + \gamma y_3^{4/3} = 1\}.$$

It is here that $c \neq 0$ plays a basic role! The choice (i) ensures Γ is an ellipsoid enclosing the origin. From the homogeneity of g, if we can show g is negative on Γ, then g will be negative definite. For $y \in \Gamma$, $|y_3| \leqq (1/\gamma)^{3/4}$, $|y_2| \leqq (1/\beta_1 c)^{1/4}$, $|y_1| \leqq 1/(4cE_1 + \beta_2 + 4E_1)^{1/4}$.

For $e > 0$ sufficiently small, $\gamma > 0$ can be made sufficiently large so that terms in $g(y)$ having a factor y_3 become negligible. Thus we have $g < 0$ on Γ if we can choose β_1, β_2, E_1, E_3 such that for $y \in \Gamma$,

$$\begin{aligned} f(y) = {} & (4E_2 - \beta_1 - c\beta_1)y_1 y_2^3 + 3\beta_1 y_1^2 y_2^2 - (\beta_2 + 4c)y_1^3 y_2 \\ & + \left(\frac{4}{3}\right) E_3 \left(\frac{1}{\beta_2}\right)^{1/3} (4E_1 y_1^3 + \beta_1 y_2^3)^{1/3} \left[\left(\frac{c}{\beta_2}\right)(4E_1 y_1^3 + \beta_1 y_2^3) + y_1^3\right] < 1. \end{aligned} \tag{27}$$

We do this by showing that the constants can be chosen so that the inequality holds where any line $y_2 = \alpha y_1$, α real, intersects Γ. Such a line intersects Γ in two points; however, from the homogeneity of f, the values of f at these points are the same. Abusing notation, let $f(\alpha)$ denote the value of f at the points where the line $y_2 = \alpha y_1$ intersects Γ, and $y_1(\alpha)$ denote the y_1 coordinate of such an intersection. Then $y_1(\alpha) \to 0$ as $\alpha \to \infty$ and the line $x_1 = 0$, or $\alpha = \infty$, must be handled separately. Here $|y_2| = (1/c\beta_1)^{1/4}$, $y_1 = 0$ and (27) requires

$$\left(\frac{4}{3}\right) cE_3 \left(\frac{1}{\beta_2}\right)^{4/3} (\beta_1)^{4/3} \left(\frac{1}{c\beta_1}\right) = \left(\frac{4}{3}\right) E_3 \left(\frac{1}{\beta_2}\right)^{4/3} \beta_1^{1/3} < 1.$$

Choose

(ii) $\beta_1^{1/3} = (\frac{3}{8})(\beta_2^{4/3}/E_3) > 0$. This choice ensures $\beta_1 > 0$ and that $\lim_{\alpha \to \infty} f(\alpha) = \frac{1}{2} < 1$ independently of future choices of β_2, E_3. Furthermore, by continuity of f, there exists an $M > 0$ such that $f(\alpha) < 1$ for $|\alpha| > M$; hence, we need only concern ourselves with $|\alpha| \leqq M$. From (27),

$$\begin{aligned} f(\alpha) = \Big\{ & (4E_2 - \beta_1 - c\beta_1)\alpha^3 + 3\beta_1 \alpha^2 - (\beta_2 + 4c)\alpha \\ & + \left(\frac{4}{3}\right) E_3 \left(\frac{1}{\beta_2}\right)^{1/3} (4E_1 + \alpha^3 \beta_1)^{1/3} \left[\left(\frac{c}{\beta_2}\right)(4E_1 + \alpha^3 \beta_1) + 1\right]\Big\} y_1^4(\alpha). \end{aligned} \tag{28}$$

Choose $E_1 = 1$ and $\beta_2 = -(4c + 3) < 0$ which satisfies (i). Next, noting (ii), choose $E_3 > 0$ large enough to make $\beta_1 < \min\{3/cM^3, 4/M^3, 1\}$. Then for $|\alpha| \leqq M$, $(4E_1 + \alpha^3\beta_1)^{1/3} > 0$, $[(c/\beta_2)(4E_1 + \alpha^3\beta_1) + 1] > 0$ and hence by increasing E_3 further, if necessary, we can assure the coefficient of $y_1^4(\alpha)$ is negative in (28), i.e., $f(\alpha) \leqq 0$ if $|\alpha| \leqq M$, and also that (25) holds, i.e., V is positive definite. Then if $u^*(x) = -(V_{x_1}(x)/4e)^{1/3}$ is the control used in system (23), the trajectory derivative of V along solutions of (23) is $\dot{V}(x) = g(x)$ which is negative definite and $x = 0$ is a globally asymptotically stable solution.

3. Concluding remarks. Our eventual goal is to show that if system (1) is STLC and satisfies additional necessary conditions, it does admit a continuous locally asymptotically stabilizing feedback control. We would also like to construct such a control. The basic approach here, shown feasible for $n = 2$, and for special cases when

$n=3$, is based on the following ideas. For the general system (1), let $\mathcal{S}^1=\{(ad^\nu X, Y)\colon \nu=0,\cdots\}$ and $\mathcal{S}^j$ denote the set of all Lie products of i-tuples of elements from $\mathcal{S}^1$ having $i\leqq j$. Let $L(\mathcal{S}^1)$ denote the Lie algebra of vector fields generated by $\mathcal{S}^1$; $L(\mathcal{S}^1)(0)$ the elements of $L(\mathcal{S}^1)$ evaluated at zero and assume $\dim L(\mathcal{S}^1)(0)=n$ which is a necessary condition for STLC. For ease of discussion, also assume the system is "odd," i.e., $\dim \operatorname{span} \mathcal{S}^{j+1}(0)=\dim \operatorname{span} \mathcal{S}^j(0)$ for odd j. This is a sufficient condition for STLC [4] and determines a dilation $\delta^r_\varepsilon x=(\varepsilon^{r_1}x_1,\cdots,\varepsilon^{r_n}x_n)$ where $r_i=1$ for $1\leqq i\leqq n_1=\dim \operatorname{span} \mathcal{S}^1(0)$, $r_i=2$ for $n_1+1\leqq i\leqq n_2=\dim \operatorname{span} \mathcal{S}^2(0)$, etc. Note that "odd" implies that all r_i are odd. Let $k=r_n$; i.e., k is the smallest integer such that $\dim \operatorname{span} \mathcal{S}^k(0)=n$. We can choose local coordinates $x=(x_1,\cdots,x_n)$, (see [6]) so that $Y=\partial/\partial x_1$ and $X(x)=X^0(x)+X^{-1}(x)+\cdots$, where X^j is homogeneous of degree j with respect to δ^r_ε. The approximating system then is $\dot{x}=X^0(x)+uY$. First use some feedback (as in the proof of Theorem 1) to eliminate the first component of X^0; i.e., we may assume $X^0(x)=\sum_{i=2}^n a_i(x)\partial/\partial x_i$, $a_i(0)=0$, $i=2,\cdots,n$. The choice of cost functional of the form

$$C(u)=\int_0^\infty [eu^{k+1}(s)+h^+_{k+1}(x(s))]\,ds, \qquad e>0,$$

with h^+_{k+1} positive definite and homogeneous of degree $(k+1)$, forces the optimal control $u^*(x,p)$ to be such that $H^*(x,p)=H(x,p,u^*(x,p))$ is homogeneous of degree $(k+1)$ with respect to the dilation $\gamma^r_\varepsilon(x,V,p)=(\delta^r_\varepsilon x, \varepsilon V, \delta^{r*}_\varepsilon p)$. Then STLC implies that the associated HJB "inequality," e.g., see (16), $-\gamma(V_{x_1}(x))^{(k+1)/k}+\sum_{j=2}^n a_j(x)V_{x_j}\leqq 0$, with equality only for $x=0$, does have a continuous viscosity solution V which is positive definite and homogeneous of degree $(k+1)$ with respect to δ^r_ε. The problem is to find additional necessary conditions for system (1), which, if satisfied, are sufficient to ensure that this solution is C^1. The homogeneity of V is such that the vector field $W(x)=X^0(x)+u^*(x, V_x(x))Y$ is homogeneous to degree zero. This means the approximating system is "good" in the sense that asymptotic stability of its zero solution implies, by Theorem 2, local asymptotic stability of the zero solution of the original system with control $u(x)=u^*(x, V_x(x))$.

Initially, the hope was that STLC and the necessary condition (B3) would be sufficient for the existence of a continuous asymptotically stabilizing feedback control for system (1). In a recent paper Coron [16] has given a necessary condition which is stronger than (B3) (i.e., it implies (B3) but the converse is not true); however, STLC and the Coron condition are not sufficient for the existence of a continuous asymptotically stabilizing feedback control even in the case when X is homogeneous of degree zero with respect to an "odd" dilation (see [17]).

REFERENCES

[1] R. W. BROCKETT, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, Vol. 27, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhauser, Boston, 1983, pp. 181–191.

[2] E. SONTAG AND H. SUSSMANN, *Remarks on continuous feedback*, in Proc. 19th Annual IEEE Conference on Decision and Control, Vol. 2, Albuquerque, New Mexico, IEEE Computer Society, Washington, DC, 1950, pp. 916–921.

[3] M. KAWSKI, *Stabilization of nonlinear systems in the plane*, Systems Control Lett., 12 (1989), pp. 169–175.

[4] H. J. SUSSMANN, *A general theorem on local controllability*, SIAM J. Control Optim., 25 (1987), pp. 158–194.

[5] H. HERMES AND M. KAWSKI, *Local controllability of a single-input affine system*, Nonlinear Analysis and Applications, Lecture Notes in Pure and Appl. Math., Vol. 109, V. Lakshmikantham, ed., Marcel Dekker, New York, 1987, pp. 235–248.

[6] G. STEFANI, *Polynomial approximations to control systems and local controllability*, in Proc. 24th Annual IEEE Conference on Decision and Control, IEEE Computer Society, Washington, DC, 1985, pp. 33-38.

[7] J. ZABZCYK, *Some comments on stabilizability*, Appl. Math. Optim., 19 (1989), pp. 1-10.

[8] W. P. DAYAWANSA, C. F. MARTIN, AND G. KNOWLES, *Asymptotic stabilization of a class of smooth two-dimensional systems*, SIAM J. Control Optim., 28 (1990), pp. 1321-1349.

[9] W. M. BOOTHBY AND R. MARINO, *Feedback stabilization of planar nonlinear systems*, Systems Control Lett., 12 (1989), pp. 87-92.

[10] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana (2), 5 (1960), pp. 102-119.

[11] E. B. LEE AND L. MARKUS, *Foundations of Optimal Control Theory*, John Wiley, New York, 1967.

[12] H. HERMES, *Controlled stability*, Ann. Mat. Pura Appl., 144 (1977), pp. 103-119.

[13] ———, *Homogeneous coordinates and continuous asymptotically stabilizing feedback controls*, International Conference on Differential Equations and Applications to Stability and Control, S. Elaydi, ed., Marcel Dekker, New York, to appear.

[14] L. C. EVANS AND M. R. JAMES, *The Hamilton-Jacobi-Bellman equation for time-optimal control*, SIAM J. Control Optim., 27 (1989), pp. 1477-1489.

[15] A. F. FILIPPOV, *Differential equations with discontinuous right hand sides*, Trans. Amer. Math. Soc., Ser. 2, 42 (1964), pp. 199-231.

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

[17] M. KAWSKI, *Homogeneous feedback laws in dimension three*, in Proc. 28th Annual IEEE Conference on Decision and Control, Tampa, Florida (1989).

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 1, pp. 197-208, January 1991

RAPID BOUNDARY STABILIZATION OF THE WAVE EQUATION*

VILMOS KOMORNIK†

Abstract. The boundary stabilization of the wave equation in bounded domains is studied. It is shown that a particular choice of the feedback leads to fast energy decay. Explicit decay rate estimates are obtained.

Key words. wave equation, boundary feedback

AMS(MOS) subject classifications. primary 35L05; secondary 93D15

1. Introduction. The decay of the local energy of solutions of the wave equation in exterior domains has been thoroughly investigated since the early 1960's; cf. Lax et al. [18], Lax and Phillips [19], Morawetz et al. [22] and their bibliography. The study of the analogous problem in bounded domains with an "energy absorbing" boundary began in the mid-1970s; cf. Quinn and Russell [23], Rauch and Taylor [24], Slemrod [27]. For the latter problem exponential energy decay was first proved by Chen [3] by adapting the multiplier techniques developed earlier for the exterior problem.

The research has been continued along two, somewhat opposite, lines:

- Finding large classes of feedbacks giving exponential decay;
- Finding special feedbacks giving fast energy decay.

In the first direction the most general results up to now have been proved by Bardos et al., [1]. In a large, natural class of feedbacks the authors have characterized those giving exponential decay. We remark that their method does not provide explicit decay rate estimates.

In the present paper we are interested in the second problem. We are going to show that a suitable special choice of the feedback leads to rapid, in some sense optimal energy decay. Moreover, strong explicit decay rate estimates will be obtained.

We turn to the formulation of our results. Let Ω be a bounded domain in $\mathbb{R}^n$ $(n \geqq 1)$ with a boundary Γ of class C^2; let us denote by ν the outward unit normal to Γ. Fix a point $x^0 \in \mathbb{R}^n$ arbitrarily and set

$$m(x) = x - x^0 \quad (x \in \mathbb{R}^n) \quad \text{and} \quad R = \sup\{|m(x)| : x \in \Omega\}. \tag{1.1}$$

Let Γ_+ and Γ_- be two disjoint open subsets of Γ such that

$$m \cdot \nu \geqq 0 \quad \text{on } \Gamma_+, \quad m \cdot \nu \leqq 0 \quad \text{on } \Gamma_-, \quad \text{and} \quad \bar{\Gamma}_+ \cup \bar{\Gamma}_- = \Gamma \tag{1.2}$$

($\cdot$ denotes the scalar product of $\mathbb{R}^n$).

Fix two nonnegative functions K, $L \in L^\infty(\Gamma_+)$ and consider the following system:

$$u'' - \Delta u = 0 \quad \text{in } \Omega \times (0, \infty), \tag{1.3}$$

$$u = 0 \quad \text{on } \Gamma_- \times (0, \infty), \tag{1.4}$$

$$\partial_\nu u + Ku' + Lu = 0 \quad \text{on } \Gamma_+ \times (0, \infty), \tag{1.5}$$

$$u(0) = u^0 \quad \text{and} \quad u'(0) = u^1 \quad \text{on } \Omega. \tag{1.6}$$

(Here and in the following ∂_ν denotes the normal derivative and $'$ the time derivative; the Laplacian Δ is taken in the space variables.) This system is well posed in the

* Received by the editors August 3, 1989; accepted for publication (in revised form) December 26, 1989. The results of this paper were obtained during a visit to Georgetown University, Washington, D.C., and Virginia Polytechnic Institute and State University, Blacksburg, Virginia.

† Université de Bordeaux I, U.F.R. de Mathématiques et d'Informatiques, 351, cours de la Liberation, 33405 Talence Cedex, France and Eötvös Loránd University, Department of Analysis, H-1088 Budapest, Múzeum Krt. 6-8, Hungary.

following sense: if we put

$$V=\{v\in H^1(\Omega)\,|\,v=0 \text{ on } \Gamma_-\}, \tag{1.7}$$

for every initial data $(u^0, u^1)\in V\times L^2(\Omega)$ the system (1.3)-(1.6) has a unique solution satisfying

$$u\in C([0,\infty);\, V)\cap C^1([0,\infty);\, L^2(\Omega)). \tag{1.8}$$

Furthermore, the system is dissipative: the energy of the solutions, defined by

$$\begin{aligned} E &= E(t)=E(u,u',t)\\ &:=\frac{1}{2}\int_\Omega u'(t)^2+|\nabla u(t)|^2\,dx+\frac{1}{2}\int_{\Gamma_+} Lu(t)^2\,d\Gamma, \end{aligned} \tag{1.9}$$

is decreasing in $t\in[0,\infty)$.

The purpose of this paper is to show that much stronger results hold if we choose the functions K and L appropriately. The main result concerns the case $n>1$.

THEOREM 1. *Fix a positive function $k\in L^\infty(\Gamma_+)$ such that*

$$k\geqq 1/R \quad \textit{and} \quad k|m|\leqq 1 \quad \textit{on } \Gamma_+ \tag{1.10}$$

and put

$$K=(m\cdot\nu)k \quad \textit{and} \quad L=\frac{n-1}{2}(m\cdot\nu)k^2. \tag{1.11}$$

(a) *For $n=2,3$ the solutions of* (1.3)-(1.6) *satisfy the following estimates*:

$$E(t)\leqq e^{1-(t/4R)}E(0) \quad \forall t\geqq 4R \quad \textit{if } n=2, \tag{1.12}$$

$$E(t)\leqq e^{1-(t/2R)}E(0) \quad \forall t\geqq 2R \quad \textit{if } n=3. \tag{1.13}$$

(b) *Let $n>3$ and assume that*

$$\bar\Gamma_+\cap\bar\Gamma_-=\varnothing. \tag{1.14}$$

Then the solutions of (1.3)-(1.6) *satisfy estimates* (1.13).

The simplest choice of a function k satisfying (1.10) is the constant function $1/R$.

In the (much simpler) case $n=1$ stronger results hold.

THEOREM 2. *Assume that $n=1$ and choose*

$$K=1 \quad \textit{and} \quad L=0. \tag{1.15}$$

Let us denote by $|\Omega|$ the length of the interval Ω. Then the solutions of (1.3)-(1.6) *have the following properties*:

$$E(t)=0 \quad \forall t\geqq|\Omega| \quad \text{if } \Gamma_-=\varnothing, \tag{1.16}$$

$$E(t)=0 \quad \forall t\geqq 2|\Omega| \quad \text{if } \Gamma_-\neq\varnothing. \tag{1.17}$$

Theorem 1 improves various earlier results proved by Chen [3], [4], Lagnese [14], [15], Quinn and Russell [23], Rauch and Taylor [24], Russell [26], Slemrod [27], Triggiani [28], Zuazua [29], [30] and Komornik and Zuazua [12], [13]; cf. also Lions [20]. Our proof will refine a method introduced in [12] and [13]; we will also use several ideas of Grisvard [5], Haraux [6], Lagnese [15], Zuazua [30], and Komornik [7]. Our results were announced in [11].

Several remarks are in order.

Remark 1.1. The geometrical hypothesis (1.14) is satisfied if Ω is star-shaped with respect to x^0 or more generally, if $\Omega=\Omega_1\backslash\bar\Omega_2$, where Ω_1,Ω_2 are star-shaped domains with respect to x^0 and $\bar{\Omega}_2 \subset \Omega_1$. It seems reasonable (but it has not yet been proved) that estimates (1.13) remain valid for $n > 3$ without hypothesis (1.14) (see Lemmas 2.2 and 4.1 below).

Remark 1.2. Theorem 1 remains valid for every bounded convex domain Ω, even if its boundary is not of class C^2, provided that $\Gamma_- = \varnothing$ (i.e., $x^0 \in \bar{\Omega}$).

Remark 1.3. If $n = 2$ and $\Gamma_- \neq \varnothing$, then the proof given below shows that, in fact, estimates (1.12) remain valid by replacing $4R$ by $4R(c+1)/(c+2)$, where c is the least positive constant satisfying

$$\int_{\Gamma_+} Lv^2 \, d\Gamma \leqq c \int_\Omega |\nabla v|^2 \, dx \quad \forall v \in V. \tag{1.18}$$

Remark 1.4. The role of the factor $(m \cdot \nu)$ in the feedback was first observed in [12] and [13], where feedbacks of the form $\partial_\nu u + k(m \cdot \nu)u' = 0$ were applied. The more general form (1.5) with $L \neq 0$ was introduced by Zuazua [29], [30] to make the feedback more robust. In Theorem 1 the special choice (1.10), (1.11) of the coefficients is essential.

Remark 1.5. According to a general principle due to Russell [26], for a system invariant with respect to time reversal, uniform exponential stabilizability implies exact controllability. In particular, estimates (1.13) yield the following exact controllability result:

Let $T > 2R$. Then for every initial data $(u^0, u^1) \in V \times L^2(\Omega)$ there exists a function $v \in L^2(0, \infty; L^2(\Gamma_+))$ such that $v(t) = 0$ for all $t > T$ and such that the solution of (1.3), (1.4), (1.6), and

$$\partial_\nu u = v \quad \text{on } \Gamma_+ \times (0, \infty) \tag{1.19}$$

satisfies $u(t) = 0$ for all $t > T$.

In other words, the "control" v drives the system to rest in time T.

At the same time, estimates (1.13) yield another exact controllability result (under the same condition $T > 2R$ but with different function spaces) concerning the system (1.3), (1.4), (1.6), and

$$u = v \quad \text{on } \Gamma_+ \times (0, \infty). \tag{1.20}$$

These results were first obtained (for every $n \geqq 1$) by Lions [20], [21], applying a direct approach, the Hilbert Uniqueness Method (HUM). (Two different, constructive versions of his proofs were later given in [7]-[10].)

If Ω is an n-dimensional ball or cube of centre x^0 (or, more generally, if Ω is contained in a ball of centre x^0 and contains at least one diameter of this ball), then in the above exact controllability theorems the number $2R$ cannot be replaced by any smaller constant. In this sense estimates (1.13) are optimal.

On the other hand, in general the constant $2R$ is not optimal; let us denote by T_0 the least constant such that the above systems are exactly controllable whenever $T > T_0$. The constant T_0 is the same for both systems; cf. [1], where T_0 was determined geometrically. It is an open question whether estimates (1.13) remain valid if we replace $2R$ by T_0.

Remark 1.6. The proof of Theorem 1, given below, may easily be generalized [17] for more general partitions (Γ_+, Γ_-) of the boundary by using, as in [14], [15], more general vector fields instead of $m(x)$. The results are, however, less explicit.

The method may also be adapted, following [13] or [16], respectively, for some nonlinear systems and for other, not necessarily hyperbolic systems.

The plan of this paper is the following. In the next section we discuss briefly the interpretation and the well-posedness of the system (1.3)-(1.6) and establish two related basic identities. Using these identities, in §3 we prove Theorem 1 in the case $n>3$ and obtain some weaker results if $n\leqq 3$. Finally, the proofs of Theorems 1 and 2 are completed for $n\leqq 3$ in §4.

2. Well-posedness and some identities. The results of this section hold for any nonnegative functions K, $L\in L^\infty(\Gamma_+)$ in (1.5); the special choice (1.10), (1.11) will not be needed here.

Assume first that system (1.3)-(1.6) has a sufficiently smooth solution. Then for every $t\in[0,\infty)$ we have $u(t)$, $u'(t)\in V$, and, multiplying (1.5) by an arbitrary function $v\in V$, integrating on Γ_+, and applying Green's formula, we find

$$\int_\Omega (\Delta u)v+\nabla u\cdot\nabla v\,dx+\int_{\Gamma_+}(Ku'+Lu)v\,d\Gamma=0. \tag{2.1}$$

This leads to the interpretation of (1.3)-(1.6) in the operational form

$$(u,u')'+A(u,u')=0 \quad \text{in} \quad (0,\infty), \tag{2.2}$$

$$(u,u')(0)=(u^0,u^1), \tag{2.3}$$

where A is a linear operator in $V\times L^2(\Omega)$ defined by

$$D(A)=\Big\{(u,z)\in V\times V\,|\,\Delta u\in L^2(\Omega),$$
$$\int_\Omega(\Delta u)v+\nabla u\cdot\nabla v\,dx+\int_{\Gamma_+}(Kz+Lu)v\,d\Gamma=0, \forall v\in V\Big\}, \tag{2.4}$$

and

$$A(u,z)=(-z,-\Delta u). \tag{2.5}$$

(We refer to, e.g., Lagnese [16] for the explanation of this interpretation in similar problems.)

The form (1.9) of the energy suggests to consider in V the euclidean norm defined by

$$\|v\|_V^2=\int_\Omega|\nabla v|^2+v^2\,dx+\int_{\Gamma_+}Lv^2\,d\Gamma,\quad v\in V; \tag{2.6}$$

it is equivalent to the norm induced by $H^1(\Omega)$.

We claim that $A+I$ is a maximal monotone operator in the Hilbert space $V\times L^2(\Omega)$. The monotonicity follows easily from (2.5), (2.6), (2.1), and from $K, L\geqq 0$: given $(u,z)\in D(A)$ arbitrarily, we have

$$\begin{aligned}
&((A+I)(u,z),(u,z))_{V\times L^2(\Omega)}\\
&=(u-z,u)_V+(z-\Delta u,z)_{L^2(\Omega)}\\
&=\int_\Omega\nabla(u-z)\cdot\nabla u+(u-z)u+(z-\Delta u)z\,dx+\int_{\Gamma_+}L(u-z)u\,d\Gamma\\
&=\int_\Omega|\nabla u|^2+u^2+z^2-uz\,dx+\int_{\Gamma_+}Kz^2+Lu^2\,d\Gamma\\
&\geqq 0.
\end{aligned}$$

For the maximal monotonicity of $A+I$ it is sufficient to show that $A+2I$ is surjective (cf., e.g., Brézis [2]). Given $(v, f) \in V \times L^2(\Omega)$ arbitrarily, consider the equation

$$\begin{aligned} &\int_\Omega \nabla u \cdot \nabla w + 4uw\,dx + \int_{\Gamma_+} (2K+L)uw\,d\Gamma \\ &\qquad = \int_\Omega (2v+f)w\,dx + \int_{\Gamma_+} Kvw\,d\Gamma \quad \forall w \in V. \end{aligned} \tag{2.7}$$

Applying the Lax-Milgram theorem, we find that (2.7) has a unique solution $u \in V$. Applying (2.7) with $w \in \mathscr{D}(\Omega)$, we obtain also that $\Delta u = 4u - 2v - f \in L^2(\Omega)$. Therefore, (2.7) may also be written in the form

$$\int_\Omega \nabla u \cdot \nabla w + (\Delta u)w\,dx + \int_{\Gamma_+} (K(2u-v) + Lu)w\,d\Gamma = 0.$$

Putting $z = 2u - v$, we obtain $(u, z) \in D(A)$. Finally, we have

$$(A+2I)(u, z) = (-z+2u, -\Delta u + 2z) = (v, f).$$

Now we apply the Hille-Yosida theory in the form given in, e.g., Brézis [2]. It follows that, for every initial data $(u^0, u^1) \in V \times L^2(\Omega)$, system (1.3)-(1.6) has a unique solution such that

$$(u, u') \in C([0,\infty); V \times L^2(\Omega)). \tag{2.8}$$

Furthermore, $e^{-t}\|(u, u')(t)\|_{V\times L^2(\Omega)}$ decreases as $t \to \infty$; in particular, from (1.9) and (2.6), it follows that

$$2E(t) \leqq e^{2t}\|(u^0, u^1)\|^2_{V\times L^2(\Omega)} \quad \forall t \geqq 0. \tag{2.9}$$

Moreover,

$$D(A) \text{ is dense in } V \times L^2(\Omega) \tag{2.10}$$

and for every initial data $(u^0, u^1) \in D(A)$ the solution of (1.3)-(1.6) has the following stronger regularity properties:

$$(u, u') \in C([0,\infty); D(A)) \cap C^1([0,\infty); V \times L^2(\Omega)). \tag{2.11}$$

Let us note that (1.3) and (2.11) imply that

$$\Delta u \in C([0,\infty); L^2(\Omega)); \tag{2.12}$$

however, we do not have

$$u \in C([0,\infty); H^2(\Omega)) \tag{2.13}$$

in general.

Let us now assume that the geometrical condition (1.14) is satisfied. Then we have $D(A) \subset H^2(\Omega) \times V$ with continuous imbedding: this may be proved in the same way as an analogous result in Lagnese [16, Chap. 3, §4.2.1.]. In this case we deduce from (2.11) that

$$u \in C([0,\infty); H^2(\Omega)) \cap C^1([0,\infty); V) \cap C^2([0,\infty); L^2(\Omega)). \tag{2.14}$$

It is well known (cf., e.g., [21]) that the relation $D(A) \subset H^2(\Omega) \times V$ also holds if Ω is convex and $\Gamma_- = \varnothing$, without any regularity assumption on its boundary. Therefore, the regularity properties (2.14) hold in this case, too.

In the rest of this section we will consider only solutions of (1.3)-(1.6) satisfying (2.14). All the calculations below will be justified by this regularity.

Remark 2.1. More precisely, the proof of Lemma 2.2 below will use (2.14). For the other calculations the regularity properties (2.11), (2.12) will be sufficient.

LEMMA 2.1. *Let u be a solution of* (1.3)-(1.6) *satisfying* (2.14). *Then*

$$E(S)-E(T)=\int_S^T\int_{\Gamma_+}K(u')^2\,d\Gamma\,dt \tag{2.15}$$

whenever $0\leqq S<T<\infty$.

Proof. Multiplying (1.3) by u', integrating by parts on $\Omega\times(S,T)$, and applying (1.4) and (1.5), we obtain

$$\begin{aligned}0&=\int_S^T\int_\Omega u'(u''-\Delta u)\,dx\,dt\\&=-\int_S^T\int_\Gamma u'\partial_\nu u\,d\Gamma\,dt+\left[\frac{1}{2}\int_\Omega (u')^2+|\nabla u|^2\,dx\right]_S^T\\&=\int_S^T\int_{\Gamma_+}K(u')^2\,d\Gamma\,dt+\left[\frac{1}{2}\int_\Omega (u')^2+|\nabla u|^2\,dx+\frac{1}{2}\int_{\Gamma_+}Lu^2\,d\Gamma\right]_S^T\end{aligned}$$

and the lemma follows from (1.9). □

Let us recall the following identity, essentially due to F. Rellich [25].

LEMMA 2.2. *If $v\in H^2(\Omega)$, then*

$$\int_\Omega 2(\Delta v)m\cdot\nabla v+(2-n)|\nabla v|^2\,dx=\int_\Gamma 2(\partial_\nu v)m\cdot\nabla v-(m\cdot\nu)|\nabla v|^2\,d\Gamma. \tag{2.16}$$

Proof. We apply Green's formula as follows:

$$\begin{aligned}&2\int_\Omega(\Delta v)m\cdot\nabla v\,dx\\&\quad=2\int_\Gamma(\partial_\nu v)m\cdot\nabla v\,d\Gamma-2\int_\Omega\nabla v\cdot\nabla(m\cdot\nabla v)\,dx\\&\quad=2\int_\Gamma(\partial_\nu v)m\cdot\nabla v\,d\Gamma-\int_\Omega 2|\nabla v|^2+m\cdot\nabla|\nabla v|^2\,dx\\&\quad=\int_\Gamma 2(\partial_\nu v)m\cdot\nabla v-(m\cdot\nu)|\nabla v|^2\,d\Gamma+\int_\Omega(n-2)|\nabla v|^2\,dx.\end{aligned}$$ □

Using Lemma 2.2 we finally establish our basic identity.

LEMMA 2.3. *Let u be a solution of* (1.3)-(1.6) *satisfying* (2.14). *Then*

$$\begin{aligned}&2\int_S^T E(t)\,dt-\int_S^T\int_{\Gamma_-}(m\cdot\nu)(\partial_\nu u)^2\,d\Gamma\,dt\\&\qquad+\left[\int_\Omega u'(2m\cdot\nabla u+(n-1)u)\,dx\right]_S^T\\&\quad=\int_S^T\int_{\Gamma_+}(m\cdot\nu)\{(u')^2-|\nabla u|^2\}\\&\qquad+\{Lu^2-(Ku'+Lu)(2m\cdot\nabla u+(n-1)u)\}\,d\Gamma\,dt\end{aligned}\tag{2.17}$$

whenever $0\leqq S<T<\infty$.

Proof. Following [12], [19], or [20], we multiply (1.3) by $2m\cdot\nabla u+(n-1)u$ and integrate by parts:

$$\begin{aligned}0&=\int_S^T\int_\Omega (2m\cdot\nabla u+(n-1)u)(u''-\Delta u)\,dx\,dt\\&=(1-n)\int_S^T\int_\Gamma u\partial_\nu u\,d\Gamma\,dt+\left[\int_\Omega u'(2m\cdot\nabla u+(n-1)u\,dx\right]_S^T\\&\quad+\int_S^T\int_\Omega -2(m\cdot\nabla u)\Delta u-m\cdot\nabla(u')^2\\&\quad+(n-1)(|\nabla u|^2-(u')^2)\,dx\,dt\\&=\int_S^T\int_\Gamma (1-n)u\partial_\nu u-(m\cdot\nu)(u')^2\,d\Gamma\,dt\\&\quad+\left[\int_\Omega u'(2m\cdot\nabla u+(n-1)u)\,dx\right]_S^T\\&\quad+\int_S^T\int_\Omega -2(m\cdot\nabla u)\Delta u+(u')^2+(n-1)|\nabla u|^2\,dx\,dt.\end{aligned}$$

Applying Lemma 2.2 shows that

$$\begin{aligned}&\int_S^T\int_\Omega (u')^2+|\nabla u|^2\,dx\,dt+\left[\int_\Omega u'(2m\cdot\nabla u+(n-1)u)\,dx\right]_S^T\\&=\int_S^T\int_\Gamma (m\cdot\nu)((u')^2-|\nabla u|^2)+\partial_\nu u(2m\cdot\nabla u+(n-1)u)\,d\Gamma\,dt.\end{aligned}$$

If we use (1.4), (1.5) and (1.9) then (2.17) follows. □

3. Proof of Theorem 1 if condition (1.14) is satisfied. Our proof is based on (2.17). (We recall that this identity was proved under hypothesis (1.14).) We begin by estimating its different terms. The special choice (1.10), (1.11) of the functions K, L will play a crucial role. In Lemmas 3.1–3.3 below we consider an arbitrary solution of (1.3)–(1.6), where $(u^0,u^1)\in D(A)$. We remark that these lemmas are valid for all $n\geqq 1$.

LEMMA 3.1. *We have*

$$\left|\int_\Omega u'(2m\cdot\nabla u+(n-1)u)\,dx\right|\leqq 2RE\quad\forall t\geqq 0 \tag{3.1}$$

where E is the expression in (1.9).

Proof. We proceed as in [7]. First we apply the divergence theorem as follows:

$$\begin{aligned}&\int_\Omega (2m\cdot\nabla u+(n-1)u)^2\,dx\\&\quad=\int_\Omega ((2m\cdot\nabla u)^2+(n-1)^2u^2+(2n-2)m\cdot\nabla(u^2)\,dx\\&\quad=\int_\Omega (2m\cdot\nabla u)^2+(1-n^2)u^2\,dx+(2n-2)\int_\Gamma (m\cdot\nu)u^2\,d\Gamma.\end{aligned}$$

Using (1.1) and (1.2), the last expression is majorized by

$$4R^2\int_\Omega |\nabla u|^2\,dx+(2n-2)\int_{\Gamma_+}(m\cdot\nu)u^2\,d\Gamma.$$

By this estimate, (3.1) now follows from the Cauchy-Schwarz inequality and (1.9)-(1.11):

$$\begin{aligned}&\left|\int_\Omega u'(2m\cdot\nabla u+(n-1)u)\,dx\right|\\&\leqq R\|u'\|^2_{L^2(\Omega)}+\frac{1}{4R}\|2m\cdot\nabla u+(n-1)u\|^2_{L^2(\Omega)}\\&\leqq R\left\{\int_\Omega (u')^2+|\nabla u|^2\,dx+\int_{\Gamma_+} Lu^2\,d\Gamma\right\}\\&=2RE.\end{aligned}$$

□

To estimate the right-hand side of (2.17) we generalize a method used in [13] for the case $L=0$.

LEMMA 3.2. *The integral on the right-hand side of* (2.17) *is majorized by*

$$\int_{\Gamma_+} 2RK(u')^2+\frac{3-n}{2}Lu^2\,d\Gamma.$$

Proof. Using (1.11) the function under the integral sign takes the following form:

$$(3.2)\qquad \begin{aligned}(m\cdot\nu)\Big\{&(u')^2-|\nabla u|^2+\frac{n-1}{2}k^2u^2\\&-\left(ku'+\frac{n-1}{2}k^2u\right)(2m\cdot\nabla u+(n-1)u)\Big\}.\end{aligned}$$

Applying the Cauchy-Schwarz inequality and using (1.10) we have

$$\begin{aligned}&\left|\left(ku'+\frac{n-1}{2}k^2u\right)2m\cdot\nabla u\right|\\&\leqq|m|^2\left(ku'+\frac{n-1}{2}k^2u\right)^2+|\nabla u|^2\\&\leqq(u')^2+\frac{(n-1)^2}{4}k^2u^2+(n-1)kuu'+|\nabla u|^2.\end{aligned}$$

Since $m\cdot\nu\geqq 0$ on Γ_+ by (1.2); hence (3.2) is majorized by

$$\begin{aligned}&(m\cdot\nu)\left\{2(u')^2+\left(\frac{n-1}{2}+\frac{(n-1)^2}{4}-\frac{(n-1)^2}{2}\right)k^2u^2\right\}\\&=(m\cdot\nu)\left\{2(u')^2+\frac{3-n}{2}\frac{n-1}{2}k^2u^2\right\}.\end{aligned}$$

Using (1.10) and (1.11) again, (3.2) is majorized by

$$2RK(u')^2+\frac{3-n}{2}Lu^2.$$

□

LEMMA 3.3. *We have*

$$(3.3)\qquad 2\int_S^T E(t)\,dt+\frac{n-3}{2}\int_S^T\int_{\Gamma_+} Lu^2\,d\Gamma\,dt\leqq 4RE(S)$$

whenever $0\leqq S<T<\infty$.

Proof. Estimating the last two integrals of (2.17) by Lemmas 3.1 and 3.2 we find that

$$2\int_S^T E(t)\,dt-\int_S^T\int_{\Gamma_-}(m\cdot\nu)(\partial_\nu u)^2\,d\Gamma\,dt$$
$$\leqq 2R(E(S)+E(T))+\int_S^T\int_{\Gamma_+} 2RK(u')^2+\frac{3-n}{2}Lu^2\,d\Gamma\,dt.$$

Applying Lemma 2.1 we conclude that

$$\begin{aligned}&2\int_S^T E(t)\,dt-\int_S^T\int_{\Gamma_-}(m\cdot\nu)(\partial_\nu u)^2\,d\Gamma\,dt\\ &\leqq 4RE(S)+\frac{3-n}{2}\int_S^T\int_{\Gamma_+}Lu^2\,d\Gamma\,dt.\end{aligned}\tag{3.4}$$

Since $m\cdot\nu\leqq 0$ on Γ_- by (1.2), the lemma follows. □

Now assume that $n\neq 2$. Then the second term in (3.3) is nonnegative because $L\geqq 0$ on Γ_+, and $L=0$ if $n=1$ (cf. (1.11)). Hence,

$$\int_S^T E(t)\,dt\leqq 2RE(S),\quad 0\leqq S<T<\infty.$$

Letting $T\to\infty$ we conclude that

$$\int_S^\infty E(t)\,dt\leqq 2RE(S)\quad\forall S\in[0,\infty)\quad\text{if } n\neq 2.\tag{3.5}$$

Now we apply a usual Gronwall-type argument as in, for example, [6] or [15]. Writing (3.5) in the form

$$\frac{d}{ds}\left(e^{s/2R}\int_s^\infty E(t)\,dt\right)\leqq 0\quad\forall s\geqq 0,$$

we conclude that

$$e^{s/2R}\int_s^\infty E(t)\,dt\leqq\int_0^\infty E(t)\,dt\quad\forall s\geqq 0.$$

Since $E(t)$ is decreasing and nonnegative (cf. Lemma 2.1 and (1.9)), the integral on the left-hand side is minorized by $2RE(s+2R)$. On the other hand, the integral on the right-hand side is majorized by $2RE(0)$ (cf. 3.5)). Hence

$$2RE^{(s/2R)}E(s+2R)\leqq 2RE(0)\quad\forall s\geqq 0,$$

which is equivalent to (1.13).

Now assume that $n=2$. Then the second term in (3.3) is not necessarily greater than or equal to zero, but it is always minorized by $-\int_S^T E(t)\,dt$ (cf. (1.9)). This leads to the estimate

$$\int_S^\infty E(t)\,dt\leqq 4RE(S)\quad\forall S\in[0,\infty)\tag{3.6}$$

instead of (3.5) and, replacing $2R$ by $4R$ everywhere in the above reasoning, we obtain estimates (3.12).

If $\Gamma_-\neq\varnothing$, then by (1.18) the second term of (3.3) is minorized also by $-(c/(c+1))\int_S^T E(t)$. This leads to the stronger estimates mentioned in Remark 1.3.

We have thus shown that if the geometrical condition (1.14) is satisfied and if $(u^0, u^1) \in D(A)$, then the solution of (1.3)-(1.6) satisfies (1.12) if $n=2$, and (1.13) if $n \neq 2$. In fact these estimates remain valid under the weaker condition $(u^0, u^1) \in V \times L^2(\Omega)$ on the initial data: this result follows easily by a standard density argument, based on (2.9) and (2.10).

In particular, Theorem 1 is completely proved for $n > 3$. □

4. Proof of the theorems for $n \leqq 3$. Let us first consider the cases $n = 2, 3$. We are going to show that in these cases the estimates (3.12), respectively, (3.13), remain valid even if (1.14) is not satisfied. Again, it is sufficient to consider initial data (u^0, u^1) belonging to $D(A)$. The difficulty is that now the solutions do not satisfy (2.17) in general.

On the other hand, let us observe that in the proof of estimates (3.12) and (3.13) in §3 we used only the inequality part "$\leqq$" of (2.17). (See also [12] and [13] on the same issue.) Hence for the proof of Theorem 1 it is sufficient to establish the following inequality, replacing (2.16).

LEMMA 4.1. *Assume that $n \leqq 3$ and let $(u, z) \in D(A)$. Then*

$$\begin{aligned} &\int_\Omega 2(\Delta u) m \cdot \nabla u + (2-n)|\nabla u|^2 \, dx \\ &\qquad \leqq \int_\Gamma 2(\partial_\nu u) m \cdot \nabla u - (m \cdot \nu)|\nabla u|^2 \, d\Gamma. \end{aligned} \tag{4.1}$$

Inequality (4.1) has been proved by Grisvard [5] in the special case $K = L = 0$. It is shown in [13] that his proof may be easily adapted for the more general case $K \geqq 0$, $L = 0$. In fact the same arguments may be used for the proof in the present case $K \geqq 0$, $L \geqq 0$. We omit the details.

Finally we turn to the proof of Theorem 2. Now Ω is a bounded open interval; let us denote it by (a, b).

Consider first the case $\Gamma_- = \varnothing$. The solution of (1.3)-(1.6) may be computed by the method of D'Alembert. Putting

$$U^1(s) = \int_\alpha^s u^1(r) \, dr, \quad s \in [a, b] \tag{4.2}$$

and then extending u^0, U^1 to $\mathbb{R}$ by

$$u^0(x) = u^0(b) \quad \text{and} \quad U^1(x) = U^1(b) \quad \text{if } x > b, \tag{4.3}$$

$$u^0(x) = u^0(a) \quad \text{and} \quad U^1(x) = U^1(a) \quad \text{if } x < a, \tag{4.4}$$

we obtain

$$2u(x, t) = u^0(x+t) + u^0(x-t) + U^1(x+t) - U^1(x-t). \tag{4.5}$$

If $t > b - a$, then from (4.2)-(4.5) we conclude that

$$2u(x, t) = u^0(b) + u^0(a) + U^1(b) - U^1(a),$$

independently of t and x. Hence (1.16) follows.

Now consider the case $\Gamma_- \neq \varnothing$. Since $\Gamma_+ \neq \varnothing$, there are two possibilities: either $\Gamma_- = \{a\}$ or $\Gamma_- = \{b\}$. We may assume by symmetry that $\Gamma_- = \{a\}$. If we define u^0 and U^1 on $\mathbb{R}$ by (4.2), (4.3), and

$$u^0(x) = -u^0(2a - x) \quad \text{and} \quad U^1(x) = U^1(2a - x) \quad \text{if } x < a, \tag{4.6}$$

the solution of (1.3)-(1.6) is given again by (4.5). In particular, for $t > 2(b-a)$ we conclude from (4.2), (4.3), (4.5), and (4.6) that

$$u(x, t) = 0 \quad \forall x \in [a, b], \quad \forall t > 2(b-a).$$

Hence (1.17) follows. □

Acknowledgments. The author is grateful to P. Fabrie, A. Haraux, J. Lagnese, and D. L. Russell for fruitful discussions.

REFERENCES

[1] C. Bardos, G. Lebeau, and J. Rauch, *Contrôle et stabilisation dans les problèms hyperboliques*, in J.-L. Lions, *Contrôlabilité exacte et stabilisation de systèmes distribués*, Vol. 1, *Contrôlabilité exacte*, Masson, Paris, 1988, pp. 492-537.

[2] H. Brézis, *Analyse fonctionnelle: théorie et applications*, Masson, Paris, 1983.

[3] G. Chen, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249-274.

[4] ——, *A note on the boundary stabilization of the wave equation*, SIAM J. Control Optim., 19 (1981), pp. 106-113.

[5] P. Grisvard, *Contrôlabilité exacte des solutions de l'équation des ondes en présence de singularités*, J. Math. Pures Appl., 68 (1989), pp. 215-259.

[6] A. Haraux, *Semi-groupes linéaires et équations d'évolution linéaires périodiques*, preprint 78011, Laboratoire d'Analyse Numérique, Université Pierre et Marie Curie, Paris.

[7] V. Komornik, *Contrôlabilité exacte en un temps minimal*, C. R. Acad. Sci. Paris Sér I Math., 304 (1987), pp. 223-225.

[8] ———, *Exact controllability in short time for the wave equation*, Ann. Inst. H. Poincaré Anal. Nonlinéaire, 6 (1989), pp. 153-164.

[9] ———, *Une méthode générale pour la contrôlabilité exacte en temps minimal*, C. R. Acad. Sci. Paris Sér I. Math., 307 (1987), 397-401.

[10] ——, *A new method of exact controllability in short time and applications*, Ann. Fac. Sci. Toulouse, to appear.

[11] ——, *Stabilisation frontière rapide de l'équation des ondes*, C. R. Acad. Sci. Paris Sér. I Math., 309 (1989), pp. 483-486.

[12] V. Komornik and E. Zuazua, *Stabilisation frontière de l'équation des ondes: une méthode directe*, C. R. Acad. Sci. Paris Sér I Math., 305 (1987), pp. 605-608.

[13] ———, *A direct method for the boundary stabilization of the wave equation*, J. Math. Pures Appl., 69 (1990), pp. 33-54.

[14] J. Lagnese, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163-182.

[15] ———, *Note on boundary stabilization of wave equations*, SIAM J. Control Optim., 26 (1988), pp. 1250-1256.

[16] ———, *Boundary Stabilization of Thin Plates*, SIAM Studies in Applied Mathematics 10, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1989.

[17] I. Lasiecka and R. Triggiani, *Uniform exponential decay in a bounded region with $L_2(0, T; L_2(\Sigma))$-feedback control in the Dirichlet boundary conditions*, J. Differential Equations, 66 (1987), pp. 340-390.

[18] P. D. Lax, C. S. Morawetz, and R. S. Phillips, *Exponential decay of solutions of the wave equation in the exterior of a star-shaped obstacle*, Comm. Pure Appl. Math., 16 (1963), pp. 477-486.

[19] P. D. Lax and R. S. Phillips, *Scattering Theory*, Academic Press, New York, 1967.

[20] J.-L. Lions, *Exact controllability, stabilizability, and perturbations for distributed systems*, SIAM Rev., 30 (1988), pp. 1-68.

[21] ———, *Contrôlabilité exacte et stabilisation de systèmes distribués*, Vol. 1, *Contrôlabilité exacte*, Masson, Paris, 1988.

[22] C. S. Morawetz, J. V. Ralston, and W. A. Strauss, *Decay of solution of the wave equation outside nontrapping obstacles*, Comm. Pure Appl. Math., 30 (1977), pp. 447-508.

[23] J. P. Quinn and D. L. Russell, *Asymptotic stability and energy decay rates for solutions of hyperbolic equations with boundary damping*, Proc. Roy. Soc. Edinburgh Sect. A, 77 (1977), pp. 97-127.

[24] J. Rauch and M. E. Taylor, *Exponential decay of solutions to hyperbolic equations in bounded domains*, Indiana Univ. Math. J., 24 (1974), pp. 79-86.

[25] F. RELLICH, *Darstellung der Eigenwerte von $\Delta u+\lambda u=0$ durch ein Randintegral*, Math. Z. 18 (1940), pp. 635-636.

[26] D. L. RUSSELL, *Controllability and stabilizability theory for linear partial differential equations. Recent progress and open questions*, SIAM Rev., 20 (1978), pp. 639-739.

[27] M. SLEMROD, *Stabilization of boundary control systems*, J. Differential Equations, 22 (1976), pp. 402-415.

[28] R. TRIGGIANI, *Wave equation on a bounded domain with boundary dissipation: an operator approach*, J. Math. Anal. Appl., 137 (1989), pp. 438-461.

[29] E. ZUAZUA, *Some remarks on the boundary stabilizability of the wave equation*, in Control of Boundaries and Stabilization, J. Simon, ed., Lecture Notes in Control and Inform. Sci., 125, Springer-Verlag, Berlin, New York, 1989.

[30] ———, *Robustesse du feedback de stabilisation par contrôle frontière*, C. R. Acad. Sci. Paris Sér I Math., 307 (1988), pp. 587-591.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 1, pp. 209-221, January 1991

ON MINIMUM ENERGY PROBLEMS*

G. DA PRATO†, A. J. PRITCHARD‡, AND J. ZABCZYK

Abstract. A stochastic system described by a semilinear equation with a small noise is considered. Under suitable hypotheses, the rate functionals for the family of distributions associated to the solution and the exit time and exit place of the solution are computed.

Key words. stochastic systems, optimization, semilinear equations, exit time

AMS(MOS) subject classifications. 93B05, 93C20, 93E20

1. Introduction. The paper is concerned with several deterministic optimization questions which arise in the theory of small noise distributed systems.

Let us assume that a stochastic system is described by a semilinear equation

$$dX^\varepsilon = (AX^\varepsilon + F(X^\varepsilon))dt + \sqrt{\varepsilon}\, dW, \quad X^\varepsilon(0) = a \in H, \qquad \varepsilon > 0, \tag{1.1}$$

where A and F are, respectively, linear and nonlinear parts of the drift term and W is a Wiener process with incremental covariance Q on a Hilbert space H.

The optimization problems considered in this paper are motivated by the problem of finding the rate functionals for the family of distributions $\mathscr{L}(X^\varepsilon(\cdot))$, $\varepsilon > 0$. They are also related to the problem of calculating (for a given domain) the exit time and exit place of the processes $X^\varepsilon(\cdot)$, $\varepsilon > 0$, (see [4], [5], [11], [12], [14], [15]). Here and in the sequel the distribution of a random variable ξ is denoted as $\mathscr{L}(\xi)$.

If $E_T(a, \cdot)$ is the rate functional for the family of measures $\mu_\varepsilon = \mathscr{L}(X^{x,\varepsilon}(T))$, $\varepsilon > 0$, then also of importance is the functional $E_\infty(a, b) = \inf_{T>0} E_T(a, b)$, $a, b \in H$, which is sometimes called the *quasipotential* ([4], [5]). For an appropriate choice of the initial condition, E_∞ is the rate functional for the invariant distributions (ν_ε) of the process X^ε.

Assume that a is a stable equilibrium point for the deterministic system $\dot{z} = Az + F(z)$, and let $\mathscr{D}$ be a set contained in H which is open with respect to the strong topology and contains the point a. Define

$$\tau^\varepsilon = \inf\{t \geqq 0;\ X^\varepsilon(t) \in \partial\mathscr{D}\}$$

then $\lim_{\varepsilon\downarrow 0} \ln \varepsilon\mathscr{E}(\tau^\varepsilon)$ is called the *exit rate.*

Now let $y^{a,\phi}(\cdot)$ be a solution to the following controlled equation:

$$\dot{y} = Ay + F(y) + Q^{1/2}\phi, \qquad y(0) = a$$

in which ϕ stands for a square integrable function from $[0, +\infty[$ into H.

Under fairly general conditions, (see [14]), we have

$$E_T(a, b) = \frac{1}{2}\inf\left\{\int_0^T \|\phi(s)\|^2\, ds;\ y^{a,\phi}(T) = b\right\}$$

and see [5], [12], and [14],

$$\lim_{\varepsilon\downarrow 0} \ln \varepsilon\mathscr{E}(\tau^\varepsilon) = \inf_{b\in\partial\mathscr{D}} E_\infty(a, b). \tag{1.2}$$

* Received by the editors September 8, 1989; accepted for publication February 7, 1990.

† Scuola Normale Superiore, Pisa, Italy.

‡ Institute of Mathematics, Warwick University, England.

§ Institute of Mathematics, Polish Academy of Sciences, Warsaw, Poland.

The set of all those points where the infimum in (1.2) is attained is the *exit set*, and this also has an important probabilistic interpretation (see [4], [5], [13]).

The present paper is concerned with the problem of finding or estimating E_T and E_∞ and is also concerned with the problem of minimizing E_∞ over the boundary of a given set D. We will refer to them as the minimum energy problem and the exit problem, respectively. We show that in certain important cases explicit solutions are possible.

The paper is divided into three parts. The first part is devoted to the minimum energy problem for linear systems. Here we gather partially known results. Basic formulae and estimates are given in Theorem 2.2. The next section starts from an upper estimate for the energy E_∞ which, however, is only valid locally. The main result of the paper is formulated in Theorem 3.7 which gives explicit formulae for E_∞ for the so called *gradient systems*. The first part of the theorem is an extension of a result by Friedlin [5] which also allows for a much larger class of drift terms. The second part is concerned with systems of second order in time, which are not discussed in [5]. All the basic steps of the proof are the same as those for the related results in finite dimensional spaces (see [4]); they require more sophisticated control theoretic and analytical developments. The final part presents a complete solution of the exit problem when the dynamics are linear.

This paper is a shortened version of the report [3], to which we will refer for additional details.

2. Minimum energy problem for linear systems. Consider a linear control system

$$\dot{y} = Ay + Bu, \qquad y(0) = a \in H \tag{2.1}$$

on a Hilbert space H. The operator A generates a C_0-semigroup of linear operators $S(t)$, $t \geqq 0$ and B is a bounded linear operator from a Hilbert space U into H. We will always assume $u(\cdot) \in L^2[0, T; U]$ for arbitrary $T > 0$.

The mild solution of (2.1) is given by

$$y(t) = S(t)a + \int_0^t S(t-s)Bu(s)\, ds, \qquad t \geqq 0. \tag{2.2}$$

Let us fix $T > 0$ and consider the following linear operator L_T acting from $L^2[0, T; U]$ into H:

$$L_T u = \int_0^T S(T-s)Bu(s)\, ds. \tag{2.3}$$

Thus

$$y(T) = S(T)a + L_T u.$$

Recall that if L is a bounded linear operator between Hilbert spaces H_1, H_2, then the value of its pseudoinverse operator L^{-1} at a point $y \in \operatorname{Im} L \subset H_2$ is characterized as the unique vector $x \in H_1$ such that

$$Lx = y, \qquad \langle x - z, x \rangle = 0 \quad \text{for all } z \in H_1, \qquad Lz = y.$$

Equivalently $x = L^{-1}y$ is the element with the smallest norm satisfying $Lx = y$.

It is clear that there exists a control $u(\cdot) \in L^2[0, T; U]$ transferring a to b in time T if and only if $b - S(T)a \in \operatorname{Im} L_T$, and it is clear that the control which achieves this and minimizes the functional $u \to \int_0^T \|u(s)\|^2\, ds$—called the *energy* functional—is

$$u = L_T^{-1}(b - S(T)a). \tag{2.4}$$

Let us recall that the system (2.1) is null controllable in time $T>0$ if an arbitrary state $b\in H$ can be transferred to 0 in time T. Moreover, the set $\mathcal{R}_T$ of all states which can be reached from 0 in time $T>0$ with controls $u(\cdot)\in L^2[0, T; H]$ is called the reachable space in time T. If $\mathcal{R}_T$ is the whole space then the system is said to be exactly controllable in time T. Finally a semigroup $S(t)$ is said to be stable if, for some positive constants M and ω we have $\|S(t)\|\leqq Me^{-\omega t}$ (see [1], [2]).

Define the linear operator

$$R_t=\int_0^t S(r)BB^*S^*(r)\,dr, \qquad t\geqq 0.$$

We have the following proposition (see [1], [2]).

Proposition 2.1. (i) *The function R_t, $t\geqq 0$, is the unique solution of the equation*

$$\text{(2.5)}\qquad \frac{d}{dt}\langle R_tx, x\rangle=2\langle R_tA^*x, x\rangle+\|B^*x\|^2, \qquad x\in D(A^*), \qquad t\geq 0; \qquad R_0=I.$$

(ii) *If A generates a stable semigroup then* $\lim_{t\to+\infty} R_t=R$ *exists and is the unique solution of the equation*

$$\text{(2.6)}\qquad 2\langle RA^*x, x\rangle+\|B^*x\|^2=0, \qquad x\in D(A^*).$$

The following theorem gives general results for the functionals $E_T(a, b)$, the minimal energy of transferring a to b in time T, and $E_\infty(a, b)$, $a, b\in H$, $T>0$. In its formulation we will use the convection that if an element x is not in the domain of an unbounded operator C we set $\|Cx\|=+\infty$.

Theorem 2.2. (i) *For arbitrary $T>0$ and $a, b\in H$:*

$$E_T(a, b)=\|R_T^{1/2})^{-1}(S(T)a-b)\|^2.$$

(ii) *If $S(t)$ is stable and the system* (2.1) *is null controllable in time $T_0>0$, then*

$$E_\infty(0, b)=\|(R^{1/2})^{-1}b\|^2 \qquad b\in H.$$

Moreover, there exists $C>0$, such that

$$\text{(2.7)}\qquad \|(R^{1/2})^{-1}b\|^2\leqq E_T(0, b)\leqq C\|(R^{1/2})^{-1}b\|^2, \qquad b\in H, \qquad T\geqq T_0.$$

Proof. The proof of (i) can be found, for instance, in [1]. To prove (ii) let us remark that the null controllability in time T_0 is equivalent to the fact that for a constant $C_1>0$ and all $x\in H$,

$$\int_0^{T_0}\|B^*S^*(r)x\|^2\,dr=\|R_{T_0}^{1/2}x\|^2\geqq C_1\|S^*(T_0)x\|^2.$$

But

$$\|R^{1/2}x\|^2=\sum_{k=0}^{\infty}\int_{kT_0}^{(k+1)T_0}\|B^*S^*(r)x\|^2\,dr$$

and for a constant $C_2>0$,

$$\int_0^{T_0}\|B^*S^*(r)x\|^2\,dr\leqq C_2\|x\|^2, \qquad x\in H.$$

Consequently, for some constants $M>0$, $\omega>0$, $C_3>0$,

$$\begin{aligned}\|R^{1/2}x\|^2 &= \int_0^{T_0} \|B^*S^*(r)x\|^2\,dr + C_2\sum_{k=1}^{\infty}\|S^*(kT_0)x\|^2\\ &\leqq \int_0^{T_0}\|B^*S^*(r)x\|^2\,dr + C_2M^2\sum_{k=0}^{\infty} e^{-\omega kT_0}\|S^*(T_0)x\|^2\\ &\leqq \left\{1+\frac{C_2}{C_1}M^2(1-e^{-\omega T_0})^{-1}\right\}\int_0^{T_0}\|B^*S^*(r)x\|^2\,dr\\ &\leqq C_3\|R_{T_0}^{1/2}x\|^2, \qquad x\in H.\end{aligned}$$

Hence $\operatorname{Im} R^{1/2}\subset \operatorname{Im} R_{T_0}^{1/2}$. Since the operator $(R_{T_0}^{1/2})^{-1}R^{1/2}$ is closed and thus bounded it follows that (2.14) holds. Now $\lim_{T\uparrow\infty}R_T=R$ so (2.13) must hold as well.

We now consider two special cases. Assume that $A: D(A)\subset H\rightarrow H$ is a negative definite operator on a Hilbert space H and that $C: H\rightarrow H$ is a bounded operator. The operators A and $\mathcal{A}$,

$$\mathcal{A}=\begin{bmatrix}0 & I\\ A & C\end{bmatrix},\ D(\mathcal{A})=D(A)\times D(-A)^{1/2}$$

define C_0-semigroups on H and $\mathcal{H}=D(-A)^{1/2}\times H$. The semigroups define mild solutions of the following Cauchy problems

(2.8) $\qquad \dot{x}=Ax, \qquad x(0)\in H$

(2.9) $\qquad \ddot{x}=Ax+C\dot{x}, \qquad x(0)\in D(-A)^{1/2}, \quad \dot{x}(0)\in H.$

The controlled version of (2.8)-(2.9) are

(2.10) $\qquad \dot{y}=Ay+u, \qquad y(0)=x\in H$

(2.11) $\qquad \ddot{y}=Ay+C\dot{y}+u, \qquad y(0)\in D(-A)^{1/2}), \qquad \dot{y}(0)=v\in H.$

We have the following theorem.

THEOREM 2.3. (i) *Assume that the operator A is negative definite, then the reachable set $\mathcal{R}_T$ for the system* (2.10) *is, for all $T>0$, exactly $D((-A)^{1/2})$ and*

$$E_\infty(0,b)=2\|(-A)^{1/2}b\|^2, \qquad b\in H.$$

(ii) *If in addition the operator C is negative definite, bounded and $(-C)^{1/2}$ commutes with $(-A)^{1/2}$ then the system* (2.11) *is exactly controllable and*

$$E_\infty\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}a\\b\end{bmatrix}\right)=2\|(-C)^{1/2}(-A)^{1/2}a\|^2+2\|(-C)^{1/2}b\|^2, \qquad \begin{bmatrix}a\\b\end{bmatrix}\in\mathcal{H}.$$

Proof. For the details of the proof see [3]. The proof that (2.11) is exactly controllable is similar to the one given for the one dimensional wave equation in [2], although, of course, a more general spectral decomposition is required.

This result does not generalize to arbitrary semigroups; however for analytic semigroups the first part of (i) can be generalized. To do this we must introduce the real interpolation space $D_A(1/2,2)$. We recall that $D_A(1/2,2)$ is the set of all x in H such that there exists a function $y(\cdot)\in W^{1,2}(0,\infty;H)\cap L^2(0,\infty;D(A))$ such that $y(0)=x$ (see [7]).

THEOREM 2.4. *Suppose that A generates an analytic semigroup on H, then the reachable set for system* (2.10) *does not depend on time and is equal to $D_A(1/2,2)$. Moreover the energy norm $(E_\infty(0,\cdot))^{1/2}$ is equivalent to the norm of $D_A(1/2,2)$.*

Proof. Let $T>0$, $u(\cdot)\in L^2[0,T;H]$; then the solution of (2.10) is given by

$$y(t)=\int_0^t S(t-s)u(s)\,ds$$

and we have (see [8]) that $y(\cdot)\in W^{1,2}(0,\infty;H)\cap L^2(0,\infty;D(A))$. Hence $y(T)\in D_A(1/2,2)$.

Conversely let $x\in D_A(1/2,2)$; then we want to show that there exists a control $u(\cdot)\in L^2[0,T;H]$ such that $y(T)=x$. Now since $x\in D_A(1/2,2)$ there exists $z(\cdot)\in W^{1/2}(0,\infty;H)\cap L^2(0,\infty;D(A))$ such that $z(0)=x$. Choose a C^∞ real function $\phi(\cdot)$ such that $\phi(0)=0$, $\phi(T)=1$ and set

$$\xi(t)=\phi(t)z(T-t),\qquad u(t)=\dot\xi(t)-A\xi(t);\qquad t\in[0,T]$$

Then $u(\cdot)L^2[0,T;H]$, $\xi(t)=y(t)$ and $\xi(T)=y(T)=x$ as required.

We now consider

$$y(t)=S(t)x+\int_0^t S(t-s)u(s)\,ds$$

and first suppose that $x\in D_A(1/2,2)$. Then there exists $z(\cdot)\in W^{1,2}(0,\infty;H)\cap L^2(,\infty;D(A))$ such that $z(0)=x$. Let h be a real valued C^∞ function such that $h(0)=1$, $h(T)=0$ and set $y(t)=h(t)z(t)$, then $y(0)=x$, $y(T)=0$ and $\dot y(\cdot)-Ay(\cdot)\in L^2[0,T;H]$. Thus it suffices to choose $u(t)=\dot y(t)-Ay(t)$. For $x\in H$ we first choose $u(t)=0$ in $[0,T/2]$; thus $y(T/2)\in D_A(\frac12,2)$. Now by the previous argument we can find a control $u(\cdot)\in L^2[T/2,T;H]$ such that $y(T)=0$.

3. Minimum energy problems for nonlinear systems. Results like Theorems 2.2, 2.4 for linear systems do not have immediate generalizations to nonlinear ones. However local results can be obtained via linearization as we shall show in Theorem 3.1. This theorem will also play a role in proving Theorem 3.7, which is an extension of Theorem 2.3 and is the most important result of the paper.

Denote by V_T the space Im $L_T=$ Im $R_T^{1/2}$ associated with the control system (2.1), equipped with the norm $\|.\|_T$:

$$\|x\|_T=\|(R_T^{1/2})^{-1}x\|=\|L_T^{-1}x\|.$$

It follows immediately from the control theoretic interpretation that if $t\leqq s$,

$$V_t\subset V_s\quad\text{and}\quad \|x\|_s\leqq\|x\|_t\qquad x\in V_t.$$

Let us assume that for all $T>0$ sufficiently small, $F: V_T\to U$ and for all $r>0$, there exists $N_{r,T}>0$ such that

$$\|F(a)-F(b)\|_U\leqq N_{r,T}\|a-b\|_T\qquad\text{provided } \|a\|_T\leqq r,\quad \|b\|_T\leqq r. \tag{3.1}$$

Consider the following equation:

$$\dot y=(Ay+BF(y))+Bu,\qquad y(0)=0, \tag{3.2}$$

which has the following mild form:

$$y(t)=\int_0^t S(t-r)BF(y(r))\,dr+\int_0^t S(t-r)Bu(r)\,dr. \tag{3.3}$$

THEOREM 3.1. *If* $2N_{r,T}\sqrt{T}<1$ *and* $\|b\|_T\leqq r((1-2N_{r,T}\sqrt{T})/2N_{r,T}\sqrt{T})$ *then*

$$E_T(0,b)\leqq(\|b\|_T+rN_{r,T}\sqrt{T})^2.$$

We will need the following result, also of independent interest.

PROPOSITION 3.2. *A mild solution* $y(\cdot)$ *of* (2.1) *with initial condition* 0 *is* V_T*-continuous on* $[0,T]$.

Proof. Fix $0 \leqq t \leqq s \leqq T$, then

$$y(s)-y(t)=\int_0^s S(s-r)Bu(r)\,dr-\int_0^t S(t-r)Bu(r)\,dr$$
$$=\int_0^s S(s-r)B\{u(r)-u(r-(s-t))I_{[s-t,s]}(r)\}\,dr.$$

Thus, from the definition of the norms in V_s and V_T,

$$(3.4)\qquad \|y(s)-y(t)\|_T^2 \leqq \|y(s)-y(t)\|_s^2 \leqq \int_0^s \|u(r)-u(r-(s-t))I_{[s-t,s]}(r)\|^2\,dr$$
$$=\int_{s-t}^s \|u(r)-u(r-(s-t))\|^2\,dr+\int_{s-t}^s \|u(r)\|^2\,dr.$$

But the right-hand side of (3.4) tends to 0 as $s-t \to 0$ and so the result follows.

Proof. The equation (3.2) can be written as

$$y(t)=L_tF(y)+L_tu$$

where $F(y)$ denotes the function $F(y(s))$ $s \in [0, T]$. If there exists a control $u(\cdot)$ that transfer zero to b in time T, then

$$(3.4)\qquad x=L_TF(y)+L_Tu.$$

Set

$$(3.5)\qquad u=L_T^{-1}(b-L_TF(y)).$$

We will now show that the following equation

$$(3.6)\qquad y(t)=L_tF(y)+L_tL_T^{-1}(b-L_TF(y)) \qquad t \in [0, T]$$

has a V_T-continuous solution. Note that then necessarily

$$y(T)=L_TF(y)+x-L_TF(y)=x$$

and the transferring control is given by (4.6).

For $z \in Z=C[0, T; V_T]$ define $\phi(z)$ by

$$\phi(z)(t)=L_tF(z)+L_tL_T^{-1}(x-L_TF(z)).$$

It follows from Theorem 4.1 that ϕ: $Z \to Z$. Note

$$\phi(0)(t)=L_tL_T^{-1}x \quad t \in [0, T]$$

and hence

$$\sup_{t \leqq T} \|L_tL_T^{-1}b\|_T^2 \leqq \int_0^T \|L_T^{-1}b(s)\|_U^2\,ds=\|b\|_T^2.$$

So $\|\phi(0)\|_Z \leqq \|b\|_T$. Let $w, z \in Z$, then

$$\phi(w)(t)-\phi(z)(t)=L_t[F(w)-F(z)]+L_tL_T^{-1}(L_T[F(z)-F(w)])$$

and hence

$$\|\phi(w)-\phi(z)\|_Z \leqq \|L.(F(w)-F(z))\|_Z-\|L.L_T^{-1}(L_T[F(z)-F(w)])\|_Z$$
$$\leqq 2\left\{\int_0^T \|F(w(s))-F(z(s))\|_U^2\,ds\right\}^{1/2}.$$

If $\|w\|_Z \leqq r$, $\|z\|_Z \leqq r$, then

$$\|\phi(w)-\phi(z)\|_Z \leqq 2N_{r,T}\left\{\int_0^T \|w(s)-z(s)\|_T^2\, ds\right\}^{1/2} \leqq 2N_{r,T}\sqrt{T}\|w-z\|_Z.$$

To show that the iterates $z_n = \phi^n(0)$, $n = 1, 2, \cdots$, are convergent it is enough to prove that $\|z_n\|_Z \leqq r$, for $n = 1, 2, \cdots$. Set $k = 2N_{r,T}\sqrt{T}$, then

$$\|\phi^n(0)\|_Z \leqq \|\phi^n(0)-\phi^{n-1}(0)\|_Z + \|\phi^{n-1}(0)-\phi^{n-2}(0)\|_Z + \cdots + \|\phi^2(0)-\phi^1(0)\|_Z$$

$$\leqq (k^{n-1}+\cdots+k)\|\phi(0)\|_Z \leqq \frac{k}{1-k}\|b\|_T.$$

So by the induction argument if

$$\frac{2N_{r,T}\sqrt{T}}{(1-2N_{r,T}\sqrt{T})}\|xb\|_T \leqq r,$$

then $\|\phi^n(0)\|_Z \leqq r$ for all $n = 1, 2, \cdots$. The sequence $\{z_n\}$ is thus convergent in Z to a solution $y(\cdot)$ of the equation (3.6). Now

$$\left\{\int_0^T \|u(s)\|^2\, ds\right\}^{1/2} = \|b - L_T F(u)\|_T \leqq \|b\|_T + \left\{\int_0^T \|F(y(s))\|_U^2\, ds\right\}^{1/2}$$

$$\leqq \|b\|_T + N_{r,T}\left\{\int_0^T \|y(s)\|_T^2\, ds\right\}^{1/2} \leqq \|b\|_T + rN_{r,T}\sqrt{T}.$$

This complete the proof.

Remark 3.3. With a similar proof to the one above we can show that there exists a unique solution of equation (3.2) on the interval $[0, T]$ for any control satisfying

$$2N_{r,T}\sqrt{T} < 1 \quad \text{and} \quad \sup_{t\leqq T}\|L_t u\|_T < r\frac{1-2N_{r,T}\sqrt{T}}{2N_{r,T}\sqrt{T}}.$$

Also, nonzero initial states can be taken into account.

COROLLARY 3.4. *Assume that for a given $T>0$ the transformation F satisfies* (3.1) *with $N_{r,T}\downarrow 0$ as $r\downarrow 0$. Then for arbitrary $\varepsilon>0$ there exists $\delta>0$ such that if $\|b\|_T<\delta$, then $E_T(0,b)\leqq\varepsilon$.*

Proof. The result follows immediately from Theorem 3.1.

We will show that Theorem 2.3 can be extended to nonlinear systems of the form

$$\dot{y} = Ay - U'(y) + u, \qquad y(0) = a_0 \in H \tag{3.7}$$

$$\ddot{y} = Ay - U'(y) - \beta\dot{y} + u, \qquad y(0) = a_0 \in D(-A)^{1/2}, \qquad \dot{y}(0) = b_0 \in H. \tag{3.8}$$

We will make the following assumptions

(i) *A is a negative definite operator on the Hilbert space H.*

(ii) *U is a functional from $V = D((-A)^{1/2})$ into R_+ of class C^1, $U(0)=0$, $DU(0)=0$.*

(iii) *There exists a mapping $U'\colon V\to H$, Lipschitz on bounded sets such that*

$$DU(x;h) = \langle U'(x), h\rangle, \quad \text{for all } x, h\in V,$$

where $DU(x; h)$ denotes the value of the Fréchet derivatives at x in the direction h.

(iv) *β is a positive constant.*

Example 3.5. Let $A = (d^2/dx^2)$, $D(A) = W_0^1(0, L)\cap W^2(0, L)$. For any positive integer k, we shall denote by $W^k(0, L)$ the Sobolev space consisting of all the real functions on $[0, T]$ which have square integrable derivatives of any order less or equal to k. Moreover, we set $W_0^1(0, L) = \{u \in W^1(0, L);\ u(0) = u(L) = 0\}$. Then $V = D(-A)^{1/2} = W_0^1(0, L)$. Define

$$U(x) = \int_0^L \phi(x(s))\,ds, \qquad x \in V,$$

where ϕ is a real valued function of class C^1. It is easy to see that

$$U'(x)(s) = \phi'(x(s)), \qquad s \in [0, L], \qquad x \in V.$$

The assumptions (ii) and (iii) are satisfied in this case.

Example 3.6. The functional $U(x) = \|(-A)^{1/2}x\|^2, x \in V$ obviously satisfies the condition (ii) and $DU(x; h) = 2\langle(-A)^{1/2}x, (-A)^{1/2}h\rangle$, $x, h \in V$. But $U'(x)$ is defined only for $x \in D(A)$, so (iii) does not hold.

The minimal energy required to transfer a_0 to a_1 for the system (3.7) and from $\left[\begin{smallmatrix} a_0 \\ b_0 \end{smallmatrix}\right]$ to $\left[\begin{smallmatrix} a_1 \\ b_1 \end{smallmatrix}\right]$ for the system (3.8) will be denoted by $E_T(a_0, a_1)$ and $E_T(\left[\begin{smallmatrix} a_0 \\ b_0 \end{smallmatrix}\right], \left[\begin{smallmatrix} a_1 \\ b_1 \end{smallmatrix}\right])$, respectively. Also $E_\infty(\cdot, \cdot) = \inf_{T>0} E_T(\cdot, \cdot)$.

We denote by $z^a(\cdot)$, $z\left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right](\cdot)$ the solutions of the uncontrolled systems

$$\dot{z} = Az - U'(z), \qquad z(0) = a \in H \tag{3.9}$$

$$\ddot{z} = Az - U'(z) - \beta\dot{z}, \qquad z(0) = a \in D(-A)^{1/2}, \qquad \dot{z}(0) = b \in H. \tag{3.10}$$

THEOREM 3.7. *Assume that the assumptions* (i)-(iv) *hold.*

(1) *If* $a \notin D(-A)^{1/2}$ *then* $E(0, a) = +\infty$.

(2) *If* $a \in D(-A)^{1/2}$ *and* $(-A)^{1/2}z^a(t) \to 0$ *as* $t \to \infty$ *in* H, *then*

$$E_\infty(0, a) = \|(-A)^{1/2}a\|^2 + 2U(a). \tag{3.11}$$

(3) *If* $\left[\begin{smallmatrix} a_0 \\ b_0 \end{smallmatrix}\right] \in \mathscr{H}$ *and* $z^{\left[\begin{smallmatrix} a \\ b \end{smallmatrix}\right]}(t) \to 0$ *as* $t \to \infty$ *in* $\mathscr{H}$, *then*

$$E_\infty\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} a \\ b \end{bmatrix}\right) = \beta[\|(-A)^{1/2}a\|^2 + 2U(a) + \|b\|^2]. \tag{3.12}$$

Proof. The proof is based on the following identities. For the system (3.7), with $y(0) \in D(-A)^{1/2}$

$$\begin{aligned} \frac{1}{2}\int_0^T \|u(s)\|^2\,ds = \frac{1}{2}\int_0^T &\|u(s) + 2Ay(s) - 2U'(y(s))\|^2\,ds \\ &+ [\|(-A)^{1/2}y(T)\|^2 + 2U(y(T)) - \|(-A)^{1/2}y(0)\|^2 \\ &- 2U(y(0))]. \end{aligned} \tag{3.13}$$

For the system (3.8), with $y(0) \in D(-A)^{1/2}$, $\dot{y}(0) \in H$,

$$\begin{aligned} \frac{1}{2}\int_0^T \|u(s)\|^2\,ds = \frac{1}{2}\int_0^T &\|u(s) - 2\beta\dot{y}(s)\|^2\,ds \\ &+ \beta[\|(-A)^{1/2}y(T)\|^2 + 2U(y(T)) + \|\dot{y}(T)\|^2 \\ &- \|(-A)^{1/2}y(0)\|^2 - 2U(y(0)) - \|\dot{y}(0)\|^2]. \end{aligned} \tag{3.14}$$

To show that (3.13) holds let us use the fact that the mild solution of (3.7) is in fact a strong solution. Elementary calculations give

$$\begin{aligned} \backslash f(1, 2)\backslash i(0, T, \|u(s)\|^2\,ds) = \frac{1}{2}\int_0^T &\|u(s) + 2Ay(s) - 2U'(y(s))\|^2\,ds) \\ &- 2\int_0^T \langle\dot{y}(s), Ay(s) - U'(y(s))\rangle\,ds \end{aligned} \tag{3.15}$$

It remains to show that

$$\int_0^T \langle \dot{y}(s), U'(y(s))\rangle \, ds = U(y(T)) - U(y(0)) \tag{3.16}$$

and

$$-2\int_0^T \langle \dot{y}(s), Ay(s)\rangle \, ds = \|(-A)^{1/2}y(T)\|^2 - \|(-A)^{1/2}y(0)\|^2. \tag{3.17}$$

In fact the identities (3.16) and (3.17) are true for arbitrary functions $y(\cdot)$ from $W^{1,2}(0, T; H) \cap L^2(0, T; D(A))$. To see this consider a sequence $\{y_n(\cdot)\}$ of functions from $C^1(0, T; D(A))$ converging to y both in $W^{1,2}(0, T; H)$ and in $L^2(0, T; D(A))$ topologies. Such a sequence exists since the domain $D(A)$ is dense in H. For each n and all $t \in [0, T]$

$$\frac{d}{dt} U(y_n(t)) = DU(y_n(t); \dot{y}_n(t)) = \langle U'(y_n(t)), \dot{y}_n(t)\rangle$$

and

$$\frac{d}{dt}\|(-A)^{1/2}y_n(t)\|^2 = \frac{d}{dt}\langle Ay_n(t), y_n(t)\rangle = 2\langle Ay_n(t), \dot{y}_n(t)\rangle.$$

So the identities (3.16) and (3.17) hold for each y_n, $n = 1, 2, \cdots$. However, we can pass to the limit in the above identities and therefore (3.16) and (3.17) hold for general

To prove (3.14) note that the functional U is defined on all state space and is of class C^1. Thus if the control $u(\cdot)$ is smooth and initial condition is in the domain of the generator then

$$\begin{aligned}\frac{1}{2}\int_0^T \|u(s)\|^2 \, ds &= \frac{1}{2}\int_0^T \|u(s) - 2\beta\dot{y}(s) + 2\beta\dot{y}(s)\|^2 \, ds \\ &= \frac{1}{2}\int_0^T \|u(s) - 2\beta\dot{y}(s)\|^2 \, ds + 2\int_0^T \langle \dot{y}(s) - Ay(s) + U'(y(s)), \dot{y}\rangle \, ds\end{aligned}$$

and consequently (3.14) holds in this case. The general case is obtained by a standard approximation argument.

Note that if $\dot{z}(t) = Az(t) - U'(z(t))$, $t \in [0, T]$, then for $y(t) = z(T-t)$ we have

$$\dot{y}(t) + Ay(t) - U'(y(t)) = 0, \qquad y(0) = z(T), y(T) = z(0), \qquad t \in (0, T).$$

Moreover, the function $y(\cdot)$ is a solution of (3.7) when $u(t) = -2Ay(t) + 2U'(y(t))$, $t \in [0, T]$, and for this control the first term on the right-hand side of (3.13) vanishes. Hence $u(\cdot) \in L^2[0, T; H]$ and

$$\begin{aligned}&E_T(0, a) \geqq \|(-A)^{1/2}a\|^2 + 2U(a)\\ &E_T(z^a(T), a) = \|(-A)^{1/2}a\|^2 + 2U(a) - \|(-A)^{1/2}z^a(T)\|^2 - 2U(z^a(T)).\end{aligned} \tag{3.18}$$

But

$$E_{T+1}(0, a) \leqq E_1(0, z^a(T)) + E_T(z^a(T), a).$$

Since $\|(-A)^{1/2}z^a(T)\| \to 0$ as $T \to \infty$ it follows from Theorem 3.1 but $E_1(0, z^a(T)) \to 0$ as $T \to \infty$. Thus

$$\lim_{T\to\infty} E_T(z^a(T), a) = \|(-A)^{1/2}a\|^2 + 2U(a)$$

and

$$\inf_{T>0} E_T(0, a) \leqq \|(-A)^{1/2}a\|^2 + 2U(a)$$

and so formula (3.11) holds. Formula (3.12) can be proved in a similar way.

Remark 3.8. Assume that Q is a linear positive definite operator that commutes with A (more precisely with the spectral measure associated with A) and consider the following system:

$$\ddot{y} = Q^{-1}Ay - \tfrac{1}{2}Q^{-1}U'(y) - \tfrac{1}{2}Q\dot{y} + u$$

$$y(0) = x \in D(-A)^{1/2}, \dot{y}(0) = v \in H.$$

Then the formula (3.12) can be generalized to the following

$$E\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} a \\ b \end{bmatrix}\right) = \|(-A)^{1/2}a\|^2 + U(a) + \langle Qb, b\rangle \tag{3.19}$$

with basically the same proof.

Remark 3.9. Let $\mathscr{E}$ be a separable Banach space containing H such that the inclusion operator i: $\mathscr{H} \to \mathscr{E}$ is radonifying. This means that if $\mathscr{E}$ is a Hilbert space then i is Hilbert–Schmidt. Then there exists an invariant measure on $\mathscr{E}$ for the process

$$dX = Ydt \qquad dY = (Q^{-1}AX + \tfrac{1}{2}Q^{-1}U'(X) - \tfrac{1}{2}QY)\,dt + dW$$

and up to a multiplicative constant is of the form

$$e^{-(1/2)U(x)}\mu\begin{bmatrix} dx \\ dy \end{bmatrix}$$

where μ is a Gaussian invariant measure for the linear system

$$dX = Y\,dt \qquad dY = (Q^{-1}AX - \tfrac{1}{2}\,QY)\,dt + dW.$$

The measure μ is cylindrical on $\mathscr{H}$ with mean vector o and covariance operator

$$R = \begin{bmatrix} Q^{-1} & 0 \\ 0 & Q^{-1} \end{bmatrix}. \tag{3.20}$$

The representation (3.20) is valid provided we introduce a new but equivalent inner product $\langle \cdot, \cdot \rangle_1$ on $\mathscr{H}$

$$\left\langle \begin{bmatrix} a_1 \\ b_1 \end{bmatrix}, \begin{bmatrix} a_2 \\ b_2 \end{bmatrix} \right\rangle_1 = \langle (-Q^{-1}A)^{1/2}a_1, (-Q^{-1}A)^{1/2}\rangle + \langle b_1, b_2\rangle.$$

The proof of this result follows from [15].

4. The exit problem. For details of the stochastic exit problem see [5], [12]. Here we discuss its deterministic analogue. To fix ideas we concentrate on the system (3.7) and assume that the conditions of Theorem 3.7 are satisfied. In addition let $\mathscr{E}$ be a Banach space containing $D(-A)^{1/2}$ and such that the inclusion operator i: $\mathscr{H} \to \mathscr{E}$ is randonifying, which means that the image $i(\gamma)$ of the cylindrical Gaussian measure $N(0, I)$ has an extension to a σ-additive measure on Borel subsets of $\mathscr{E}$.

Example 4.1 (Compare [5]). Let

$$A = \frac{d^2}{dx^2}, \qquad D(A) = W_0^1(0, L) \cap W^2(0, L).$$

Then

$$D(-A)^{1/2}=W_0^1(0,L).$$

If $E=C[0,L]$, then the inclusion $i: W_0^1(0,L)\to\mathscr{E}$ is radonifying (see [6]).

Let $\mathscr{D}$ be a bounded open set in $\mathscr{E}$ containing 0. For arbitrary $\varepsilon>0$ define

$$(\partial\mathscr{D})_\varepsilon=\{x\in\mathscr{E}, \text{distance}_{\mathscr{E}}\,(x,\partial\mathscr{D})<\varepsilon\} \tag{4.1}$$

and let

$$r_\varepsilon^-=\inf\{E_\infty(0,b);\ b\in(\partial\mathscr{D})_\varepsilon\cap\bar{\mathscr{D}}\},$$
$$r_\varepsilon^+=\inf\{E_\infty(0,b);\ b\in(\partial\mathscr{D})_\varepsilon\cap\mathscr{D}^c\},$$

where $\bar{\mathscr{D}}$ and $\mathscr{D}^c$ denote, respectively, the closure and the complement of $\mathscr{D}$.

$$\hat{r}=\inf\{E_\infty(0,b);\ b\in\partial\mathscr{D}\}.$$

If

$$r^-=\lim_{\varepsilon\downarrow 0} r_\varepsilon^-,\ r^+=\lim_{\varepsilon\downarrow 0} r_\varepsilon^+$$

then $r^-\leqq\hat{r}$, $r^+\leqq\hat{r}$ and we expect that in fact $r^-=r^+=\hat{r}$. The numbers r^-, r^+, and $\hat{r}$ will be called, respectively, the lower, the upper exit rates, and the exit rate.

The following problems are of interest for both deterministic and stochastic systems.

PROBLEM 4.2. Under what conditions $r^-=r^+=\hat{r}$?

PROBLEM 4.3. Assume that $r^-=r^+=\hat{r}$. Calculate $\hat{r}$ and describe as explicitly as possible the set

$$\hat{\mathscr{E}}=\{b\in\partial\mathscr{D});\ E_\infty(0,b)=\hat{r}\} \tag{4.2}$$

which will be called the *exit set.*

For linear systems some answers to the above questions are available assuming that $\mathscr{E}=H$(see [12]); here we consider a different situation and give rather specific answers to both the problems. Namely we consider the problem

$$\inf_{u\in\mathscr{D}^c}\|Au\|_H^2,$$

where A is a closed operator on $H=L^2(\Gamma)$ with the domain $D(A)\subset C(\bar{\Gamma})=E$, $C(\bar{\Gamma})$ being the space of continuous functions on $\bar{\Gamma}\subset R^n$ and $\mathscr{D}$ is a bounded neighborhood of 0 in E. If $u\notin D(A)$, we set $\|Au\|=+\infty$. We will assume also that:

(i) The operator $G=A^{-1}$ is an integral operator with a continuous kernel $g(\cdot,\cdot)$:

$$Gv(x)=\int_\Gamma g(x,y)v(y)\,dy,\qquad x\in\Gamma,\quad v\in H.$$

(ii) The set $\mathscr{D}$ is of the following form:

$$\mathscr{D}=\{u\in E;\ -b(x)<u(x)<a(x),x\in\Gamma\}$$

where $a(\cdot)$ and $b(\cdot)$ are positive functions on $\bar{\Gamma}$.

The following result holds.

THEOREM 4.4. *Assume* (i) *and* (ii) *hold. Then* $r^-=r^+=\hat{r}$ *and*

$$\hat{r}=\inf_{x\in\Gamma}\left\{(a(x)\wedge b(x)\left(\int_\Gamma g^2(x,y)\,dy\right)^{-1}\right\}. \tag{4.3}$$

Let Γ_0 be the set of all $x \in \Gamma$ for which the infimum in (4.3) *is attained. If Γ_0 is nonempty then*

$$\hat{\mathscr{E}} = \left\{ (Gv^x)(\cdot);\ v^x(\cdot) = \frac{\hat{r}}{a(x) \wedge b(x)} g(x, \cdot), x \in \Gamma_0 \right\}.$$

Proof. Let us fix $x \in \Gamma$ and a positive number c; first we will solve the problem: $\inf\{\|Au\|^2;\ u(x) \geqq c, u \in E\}$ (if $u \notin D(A)$, $\|Au\| = +\infty$), which is equivalent to

$$\inf_{Gv(x) \geqq c} \|v\|^2 = \inf_{Gv(x) = c} \|v\|^2 \tag{4.4}$$

where $Gv(x) = \langle g(x, \cdot). v\rangle_H$. The proof of the following lemma is straightforward.

LEMMA 4.5. *Let H be a Hilbert space, $v \in H$ and $c > 0$, then for the problem $\inf\{\|u\|_H;\ \langle u, v\rangle = c\}$ the infimum is attained at $u = cv\|v\|^{-2}$ and is equal to $c\|v\|^{-1}$.*

Therefore the problem (4.4) has a unique solution $v^x(\cdot) = cg(x, \cdot)(\int_\Gamma g^2(x, y)\, dy)^{-1}$ and the minimum value is $c^2\|v^x(\cdot)\|^{-2}$. The statement of the theorem now follows easily.

Example 4.6. Let $A_1 = (d^2/dx^2)$, $D(A_1) = W_0^1(0, 1) \cap W^2(0, 1)$, $\mathscr{E} = C_0(0, 1)$ the space of continuous functions vanishing at 0 and 1, $U' = 0$ and

$$\mathscr{D} = \{z \in \mathscr{E};\ |z(x)| < a, x \in [0, 1]\}.$$

PROPOSITION 4.7. *For the above example $r^- = r^+ = \hat{r} = 4a^2$ and the exit set consists of exactly two functions $\pm\hat{z}$*

$$\hat{z}(t) = \begin{cases} (a/2)t & \text{if } t \in [0, \frac{1}{2}] \\ (a/2)(1-t) & \text{if } t \in [\frac{1}{2}, 1]. \end{cases}$$

Proof. The proof follows from Theorem 4.4 and elementary calculations.

Example 4.8. Here we take $A_2 = -A_1^2$ where A_1 is the same as in Example 4.6. Note that then $D(-A_2)^{1/2} = D(A_1)$ and

$$\|(-A_2)^{1/2}z\| = \int_0^1 \left[\frac{d^2z}{dx^2}(x)\right]^2 dx.$$

The set $\mathscr{D}$ is the same as that in Example 4.6.

PROPOSITION 4.9. *For the above example $r^- = r^+ = \hat{r} = 48a^2$ and the exit set consists of exactly two functions $\pm\hat{z}$*

$$\hat{z}(t) = \begin{cases} at(3-4t^2) & \text{if } t \in [0, \frac{1}{2}] \\ a(1-t)(3-4(1-t)^2) & \text{if } t \in [\frac{1}{2}, 1]. \end{cases}$$

Remark 4.10. It is clear that a similar result is true for the operator $A_n = (-1)^{n+1}A_1^n$. We could also start from the operator $A_1 = (d^2/dx^2)$ on $L^2(0, 1; \mathbf{R}^d)$ of square integrable vector functions. $\mathscr{E} = C(0, 1; \mathbf{R}^d)$ and the set $\mathscr{D}$ could be of more general character

$$\mathscr{D} = \{z;\ z(s) \in T(s), s \in [0, 1]\}$$

where T is a multifunction with values in $\mathbf{R}^d$. Some additional subtleties arise here.

Remark 4.11. It would be interesting to consider in detail the case $A_1 = \Delta$ on $L^2(\Gamma)$, Γ bounded in $\mathbf{R}^n$, $D(A_1) = W_0^1(\Gamma) \cap W^2(\Gamma)$ and $A_m = (-1)^{m+1}A_1^m$. Under well-known conditions, $D(A_m) \subset c(\bar{\Gamma})$ and the exit problem, as formulated in Problems 1 and 2 can be posed correctly. Some related comments can be found in [5].

REFERENCES

[1] R. F. CURTAIN, *Linear-quadratic control with fixed end points in infinite dimensions*, J. Optim. Theory Appl., 44 (1984), pp. 55–74.

[2] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sci. 8, Springer-Verlag, Berlin, New York, 1978.

[3] G. Da Prato, A. J. Pritchard, and J. Zabczyk, *On minimum energy problems*, Control Theory Center Report No. 156, Warwick University, England.

[4] M. Friedlin and A. Wentzell, *Random Perturbations of Dynamical Systems*, Springer-Verlag, Berlin, New York, 1987.

[5] M. Friedlin, *Random perturbation of reaction diffusion equations*, Trans. Amer. Math. Soc., 305 (1988), pp. 665-697.

[6] H. Kuo, *Gaussian measures in Banach Space*, Lecture Notes in Math. 463, Springer-Verlag, Berlin, New York, 1975.

[7] J. L. Lions and E. Magenes, *Problèmes aux Limites Non Homogènes et Applications*, Dunod, Paris, 1968.

[8] J. L. Lions and J. Peetre, *Sur une classe d'espaces d'interpolation*, Inst. Hautes Etudes Scientifiques, 19 (1964), pp. 5-68.

[9] K. Magnusson, A. J. Pritchard, and M. D. Quinn, *The application of fixed point theorems to global nonlinear controllability problems*, Banach Centre Publications, 14 (1985), pp. 319-344.

[10] A. J. Pritchard and J. Zabczyk, *Stability and stabilizability of infinite dimensional systems*, SIAM Rev., 23 (1983), pp. 25-52.

[11] J. Zabczyk, *Structural properties and limit behaviour of linear stochastic systems in Hilbert spaces*, Banach Centre Publications, 14 (1985), pp. 591-609.

[12] ———, *Exit problems for infinite dimensional systems*, Lecture Notes in Math., 1236 (1987), pp. 249-296.

[13] ———, *Exit problems and control theory*, Systems Control Lett., 6 (1985), pp. 165-172.

[14] ———, *On large deviations for stochastic evolution equations*, Proceedings of the 6th IFIP Working Conference on Stochastic Systems and Optimization, Warsaw, 1988.

[15] ———, *Symmetric solutions of semilinear stochastic equations*, Lecture Notes in Math., 1390 (1989), pp. 237-256.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 1, pp. 222-252, January 1991

PSEUDODIFFERENTIAL PERTURBATIONS AND STABILIZATION OF DISTRIBUTED PARAMETER SYSTEMS: DIRICHLET FEEDBACK CONTROL PROBLEMS*

MICHAEL PEDERSEN†

Abstract. The stabilization problems for parabolic and hyperbolic partial differential equations with Dirichlet boundary condition are considered. The systems are stabilized by a boundary feedback in

(1) The operator equation,

(2) The boundary condition,

(3) Both the operator equation and the boundary condition;

the existence of feedback semigroups in these cases is also proved. The main tool in the investigation is a pseudodifferential transformation that transforms the domains of the feedback semigroup generators into classical operator domains, where a direct resolvent analysis can be employed. The transformation turns out to be a shortcut to some of the stabilization results of Lasiecka and Triggiani in [*J. Differential Equations*, 47 (1983), pp. 245-272], [*SIAM J. Control Optim.*, 21(1983), pp. 766-802], and [*Appl. Math. Optim.*, 8(1981), pp. 1-37], and it illuminates to some extent how a change of boundary condition influences the systems.

Key words. partial differential equations, stabilization, pseudodifferential operators, boundary feedback control

AMS(MOS) subject classifications. 35J05, 93D25

Orientation. This paper is concerned with boundary feedback stabilization problems for parabolic and hyperbolic evolution equations associated with elliptic differential operators. Boundary feedback systems have been studied intensively during the last 10 years by Lasiecka and Triggiani and others; we refer here only to [7]-[9]. Lasiecka and Triggiani employ a semigroup-based method, related to work of Washburn [15] and Balakrishnan [1]. This semigroup-integral representation method is based on the theory of fractional power spaces and second-order operators are considered. Lasiecka and Triggiani prove the existence of a feedback semigroup on negative order fractional power spaces strictly larger than $L^2(\Omega)$, and then the semigroup is restricted to $L^2(\Omega)$.

We describe here a pseudodifferential operator approach that allows us to work directly in $L^2(\Omega)$, and we construct in a direct way the resolvent of the 2m-order operator of the evolution equation we consider. The explicit resolvent construction is then used to derive stabilization results for the boundary feedback semigroup of the problem, and to improve some of the estimates given in [7]-[9].

Another advantage is the application of the pseudodifferential point of view to *characterize* various types of feedback systems appearing in the literature. Some of the first results on the stabilization of feedback systems are due to Nambu [11] and Triggiani [13], where the systems are stabilized by a suitable manipulation (control) in the operator equation. Such manipulations are included in what we denote *perturbations of the first kind.* Another possibility, usually considered more complicated than manipulating with the operator equation, is the manipulation with the boundary condition. Stabilization of a system by the changes of the boundary condition is treated in [7]-[9]. It is this kind of manipulation of the boundary condition that is usually understood

* Received by the editors January 15, 1989; accepted for publication (in revised form) January 29, 1990. This work was partially supported by the Carlsberg Foundation.

† Institute of Mathematics and Physics, Roskilde University Centre, Post Box 260, DK-4000, Roskilde, Denmark. Present address, Mathematical Institute, The Technical University of Denmark, DK 2800-Lyngby—Denmark.

as a boundary control, or as a boundary feedback for systems where the control is a feedback of the state. We denote this kind of control *a perturbation of the second kind.* For both types of perturbations it is most realistic to assume that the feedback controls are of finite rank.

One of the main results of this paper is that we can, in general, *transform* a perturbation of the second kind into the more simple and well-understood perturbation of the first kind, where a "classical" approach can be taken for obtaining stabilization results. It turns out that the transformation can be regarded as a generalized change of coordinates in the space $H^m(\Omega)$, a feature that is also of interest in the case of optimal control of systems. But the optimal problem is not elaborated on in this paper.

For the sake of generality, we also introduce a combination of the perturbations. Here we manipulate *both* the operator equation *and* the boundary condition and call this a *perturbation of the third kind.* For such systems it is possible to construct a system operator that is *variational* (i.e., associated with a suitable semibounded sesquilinear form), and calculations on this operator demonstrate the linking of "interior terms" and "boundary terms." The stabilization procedure suggested for such systems with mixed feedbacks is perhaps not optimal with respect to dimensionality, but it is simple, and the theory elucidates the nature of the boundary feedback systems.

The pseudodifferential approach thus allows us to obtain stabilization results in a unified setting for all three kinds of perturbations. Moreover, we treat parabolic as well as hyperbolic problems. In all cases we conclude that it is possible to construct finite rank feedback mechanisms that give exponential decrease of the L^2-norm of the state. For perturbations of the first and second kind this is achieved by the pole-assignment theorem, while for perturbations of the third kind positivity arguments are used.

The results on perturbations of the first and second kind are contained in § 5 of this paper, while § 6 deals with the perturbations of the third kind. Section 1 is an introduction to the notation used throughout the paper; § 2 introduces the specific form of the perturbations considered. Section 3 and 4 deal with some results of pseudodifferential calculus, most importantly the transformation techniques employed.

1. Introduction and notation. We consider stabilization of parabolic and hyperbolic differential equations of the form

$$\begin{aligned} &\partial_t u + Au = 0 && \text{in } \Omega && \text{for } t > 0,\\ &\gamma u = 0 && \text{on } \Gamma && \text{for } t > 0,\\ &u = u_0 && \text{in } \Omega && \text{for } t = 0, \end{aligned} \tag{1.1}$$

and

$$\begin{aligned} &\partial_t^2 u + Au = 0 && \text{in } \Omega && \text{for } t \in \mathbb{R},\\ &\gamma u = 0 && \text{on } \Gamma && \text{for } t \in \mathbb{R},\\ &u = u_0 && \text{in } \Omega && \text{at } t = 0,\\ &\partial_t u = u_1 && \text{in } \Omega && \text{at } t = 0. \end{aligned} \tag{1.2}$$

Here A is a formally self-adjoint, uniformly strongly elliptic differential operator of order $2m$, of the form

$$A = \sum_{|\alpha|,|\beta| \leqq m} D^\beta a_{\alpha\beta}(x) D^\alpha, \tag{1.3}$$

with $C^\infty(\bar{\Omega})$-coefficients $a_{\alpha\beta}$ on a bounded, open domain $\Omega \subset \mathbb{R}^n$, $n \geqq 2$, with smooth boundary $\partial\Omega = \Gamma$. γ is the *Dirichlet trace operator*

$$\gamma = \{\gamma_j\}_{0 \leqq j < m} \tag{1.4}$$

where

$$\gamma_j u = \left(\frac{1}{i}\frac{\partial}{\partial n}\right)^j u|_\Gamma. \tag{1.5}$$

(n is the normal, directed inward.)

We denote similarly

$$\nu = \{\gamma_j\}_{m \leqq j < 2m} \tag{1.6}$$

the *Neumann trace operator*, and we define the *Cauchy-data* ρu as

$$\rho u = \{\gamma u, \nu u\}. \tag{1.7}$$

Moreover, we use the multi-index notation

$$D^\alpha = D_1^{\alpha_1}, \cdots, D_n^{\alpha_n}, \qquad D_j^{\alpha_j} = \left(\frac{1}{i}\frac{\partial}{\partial x_j}\right)^{\alpha_j}. \tag{1.8}$$

The *Dirichlet realization* A_γ of A is the operator that acts as does A in $L^2(\Omega)$, and with domain

$$D(A_\gamma) = \{u \in H^{2m}(\Omega) \mid \gamma u = 0\} = H^{2m}(\Omega) \cap H_0^m(\Omega), \tag{1.9}$$

where $H^s(\Omega)$ is the Sobolev space of $L^2(\Omega)$-functions with $L^2(\Omega)$-derivatives up to order s. It is well known that A_γ is an unbounded, self-adjoint operator in $L^2(\Omega)$, and since the embedding $H^s(\Omega) \to H^t(\Omega)$ is compact for $s > t$, the resolvent $R(\lambda, A_\gamma)$ of A_γ is a compact operator in $L^2(\Omega)$ for all λ outside the spectrum of A_γ, $\mathrm{sp}(A_\gamma)$. Hence A_γ has a sequence of real eigenvalues $\lambda_1 \leqq \lambda_2 \leqq \cdots$ converging to infinity. We see that (1.1) and (1.2) are the time-dependent evolution problems associated with A_γ, generalizing the heat equation, respectively, the wave equation. When $\lambda_1 > 0$, all solutions $u(t, x)$ of (1.1) are exponentially decreasing for $t \to \infty$, and all solutions of (1.2) are bounded; we will call this the *stable* case. However, if some eigenvalues are negative, there are solutions both of (1.1) and (1.2) that grow in an exponential manner as $t \to \infty$. It is therefore of interest to investigate how we can change the systems to obtain the stable case, and that is the aim of this paper.

2. Perturbations of the boundary value problems. By a *perturbation of the first kind* of the system (1.1) we will understand a system of the form

$$\begin{aligned} &\partial_t u + Au + Gu = 0 && \text{in } \Omega && \text{for } t > 0, \\ &\gamma u = 0 && \text{on } \Gamma && \text{for } t > 0, \\ &u = u_0 && \text{in } \Omega && \text{at } t = 0. \end{aligned} \tag{2.1}$$

Here the interior operator A is replaced by $A + G$ where G has finite rank and is of the special form $G = KT$, where

(i) T is a *trace operator* that maps functions on Ω into functions on Γ, in the form of a column vector.

(ii) K is a *Poisson operator* that maps functions on Γ into functions on Ω, of the form of a row vector.

An operator G of this special form is denoted as a *singular Green operator.* This kind of operator is carefully explained, as well as the trace and Poisson operators, in Grubb [5]. They are all operators entering in the "Boutet de Monvel-calculus" (cf. Boutet de Monvel [2]).

Stabilization of the *a priori* unstable system (1.1) by a perturbation of the first kind has been studied, e.g., in Nambu [11] and Triggiani [14], and it is shown there that it is possible to choose G of finite rank, such that (2.1) is stable.

By a *perturbation of the second kind* of the system (1.1) we will understand a system of the form

$$\begin{aligned} &\partial_t u + Au = 0 && \text{in } \Omega && \text{for } t>0, \\ &\gamma u = T'u && \text{on } \Gamma && \text{for } t>0, \\ &u = u_0 && \text{in } \Omega && \text{at } t=0, \end{aligned} \tag{2.2}$$

where the boundary operator γ is replaced by $\gamma - T'$, T' being a trace operator of finite rank.

Both (2.1) and (2.2) are so-called *boundary feedback systems,* and we will especially be interested in the case where T' is of the special form

$$T'u = \sum_{j=1}^{N} (u \mid w_j) g_j \tag{2.3}$$

(here $(\cdot\mid\cdot)$ is the usual $L^2(\Omega)$ inner product, $w_j \in C^\infty(\bar{\Omega})$, $g_j \in C^\infty(\Gamma)^m$, $j = 1, 2, \cdots, N$).

For applications, the w_j can be thought of as sensor functions and the g_j as boundary actuators.

T' defined in (2.3) is called a *finite-dimensional feedback operator,* and in contrast to γ it is of a nonlocal nature. (One of the major differences between the perturbations (2.1) and (2.2) is that in (2.1) the boundary condition is *local,* whereas in (2.2) it is nonlocal.)

Boundary feedback systems have been studied in a number of papers by Lasiecka and Triggiani (see, e.g., Lasiecka and Triggiani [7]-[9]). One of the main results is that, under suitable hypotheses, it is possible to choose the functions w_j and g_j appearing in (2.3), such that the system is stable. Lasiecka and Triggiani took a semigroup approach to investigate the system (2.2), using developments of the semigroup approach presented in Washburn [15] and Balakrishnan [1]. (The basic idea of a semigroup model is presented in Fattorini [4], where ordinary differential equations are considered.)

By a *perturbation of the third kind* of the system (1.1) we will understand a system of the form

$$\begin{aligned} &\partial_t u + Au + Gu = 0 && \text{in } \Omega && \text{for } t>0, \\ &\gamma u = T'u && \text{on } \Gamma && \text{for } t>0, \\ &u = u_0 && \text{in } \Omega && \text{at } t=0, \end{aligned} \tag{2.4}$$

where the operators G and T' are of the types considered above.

We define perturbations of the system (1.2) in an analogous way.

In the following we present a pseudodifferential operator method to investigate the systems above. This gives us in an easy way many of the results of Lasiecka and Triggiani [7]-[9], as well as similar results for the hyperbolic problem

$$\begin{aligned} &\partial_t^2 u + Au = 0 && \text{in } \Omega && \text{for } t \in \mathbb{R}, \\ &\gamma u = T'u && \text{on } \Gamma && \text{for } t \in \mathbb{R}, \\ &u = u_0 && \text{in } \Omega && \text{at } t=0, \\ &\partial_t u = u_1 && \text{in } \Omega && \text{at } t=0 \end{aligned} \tag{2.5}$$

(a perturbation of the second kind of (1.2)).

Here we would like to point out that there now exists a quite general theory that includes all the above perturbations when they have "smooth coefficients," namely, the theory of *pseudodifferential boundary problems*. For these, the solvability of parabolic problems such as (2.2) (and far more general cases) has been discussed in great detail in Grubb [5]. However, in the work that follows, we use only some basic results of the pseudodifferential point of view. But the techniques were crucial to obtain the simplicity of the proofs, and the theory was very helpful for the understanding of the underlying problems.

One of our main results is that we can, in general, *transform* a perturbation of the second kind into a perturbation of the first kind, whenever the boundary condition is *normal*, in the sense described in Grubb [5]. This includes all "classical" normal boundary conditions, as well as the feedback boundary condition $\gamma u - T'u = 0$, with T' given by (2.3). In this case, however, a special transformation of the systems (2.2) and (2.5) proves to be very useful. It turns out that the transformation can be regarded as a generalized change of coordinates, and the resulting transformed system operator $A+G$ is merely a finite-dimensional, A-bounded perturbation of A. In this case, stabilization theory for perturbations of the first kind is straightforward, as we can apply the well-known "pole assignment theorem" due to Wonham (see Wonham [16]).

3. Normal operator realizations. Assume that $Tu = 0$ is a *normal boundary condition* in the sense of Grubb [5], i.e., the highest order normal derivatives enter with a surjective coefficient matrix. We will then define a *normal realization* of the operator A in (1.3) the following way. Let A_T be the operator that acts as does A in $L^2(\Omega)$, with domain

$$D(A_T) = \{u \in H^{2m}(\Omega) \mid Tu = 0\}. \tag{3.1}$$

Then the realization A_T of A is called a normal realization.

It is shown in Grubb [5, Lemma 1.6.8], that normal realizations have dense domains in $L^2(\Omega)$. According to Grubb, there exists an operator Λ, (see (4.4)) that is a homeomorphism in $H^s(\Omega)$ for any $s \geqq 0$, such that Λ defines a bijection

$$\Lambda\colon \quad D(A_T) \tilde{\to} D(A_\gamma) = H^{2m}(\Omega) \cap H_0^m(\Omega), \tag{3.2}$$

where A_γ is the Dirichlet realization from § 1.

Moreover, we have Lemma 3.1.

LEMMA 3.1. *Let T' be given by* (2.3) *and define*

$$T = \gamma - T'. \tag{3.3}$$

Then $Tu = 0$ is a normal boundary condition, and the operator realization A_T of A is a closed, densely defined operator in $L^2(\Omega)$.

Proof. We only have to show that A_T is closed. Let (u_n) be a sequence in $D(A_T)$ converging to $u \in L^2(\Omega)$, and assume that (Au_n) converges to $v \in L^2(\Omega)$. We must show that $u \in D(A_T)$ with $Au = v$. Since $u_n \to u$ in $L^2(\Omega)$, we have that $Au_n \to Au$ in $\mathcal{D}'(\Omega)$ (space of distributions on Ω), so that $v = Au \in L^2(\Omega)$ and $u_n \to u$ in the space $\{v \in L^2(\Omega) \mid Av \in L^2(\Omega)\}$. Now $\gamma u_n \to \gamma u$ in $\prod_{0 \leqq k < m} H^{-1/2-k}(\Gamma)$ (see Lions and Magenes [10]) and since

$$\gamma u_n = \sum_{j=1}^{N} (u_n \mid w_j) g_j \quad \text{where } (u_n \mid w_j) \to (u \mid w_j), \quad j = 1, 2, \cdots, N,$$

we see that $\gamma u_n \to \sum_{j=1}^N (u|w_j)g_j$, so γu equals $\sum_{j=1}^N (u|w_j)g_j$ as an element of $\prod_{0\leqq k<m} H^{1/2-k}(\Gamma)$. Now, since $g_j \in C^\infty(\Gamma)^m \subseteq \prod_{0\leqq k<m} H^{2m-k-1/2}(\Gamma)$, $j=1,2,\cdots,N$, we have that $\gamma u \in \prod_{0\leqq k<m} H^{2m-k-1/2}(\Gamma)$. But then by the regularity of the Dirichlet problem for A, $Au \in L^2(\Omega)$ and $\gamma u \in \prod_{0\leqq k<m} H^{2m-k-1/2}(\Gamma)$ imply that $u \in H^{2m}(\Omega)$. Altogether, $u \in H^{2m}(\Omega)$ with $\gamma u - \sum_{j=1}^N (u|w_j)g_j = 0$, so $u \in D(A_T)$. □

4. The pseudodifferential transformations. Consider for $l=1,2$ the parabolic, respectively, hyperbolic perturbation of the second kind

$$\partial_t^l u + A_T u = 0, \qquad u \in D(A_T), \tag{4.1}$$

with $T=\gamma-T'$, T' given by (2.3).

Using (3.2) this can be transformed into

$$\partial_t^l \Lambda^{-1} v + A_T \Lambda^{-1} v = 0, \qquad v \in D(A_\gamma), \tag{4.2}$$

where $v=\Lambda u$. Acting with Λ from the left in (4.2) we find

$$\partial_t^l v + \Lambda A_T \Lambda^{-1} v = 0, \qquad v \in D(A_\gamma). \tag{4.3}$$

It is shown in Lemma 1.6.8 of Grubb [5] that Λ and Λ^{-1} can be chosen in the following form:

$$\Lambda = 1 - K_0 T', \qquad \Lambda^{-1} = 1 - K_0 Q_0 T', \tag{4.4}$$

where K_0 is a certain standard type of Poisson operator, chosen such that K_0T' has small norm, a Q_0 is a certain pseudodifferential operator, that is bijective and elliptic in $\prod_{0\leqq k<2m} H^{s-k}(\Omega)$, $s\geqq 0$. Then

$$\Lambda A_T \Lambda^{-1} = (1-K_0T')A_T(1-K_0Q_0T') = A_T + G, \tag{4.5}$$

where

$$G = K_0 T' A_T K_0 Q_0 T' - A_T K_0 Q_0 T' - K_0 T' A_T, \tag{4.6}$$

is a singular Green operator of finite rank.

Hence (4.3) is a perturbation of the first kind.

We have thus obtained Proposition 4.1.

PROPOSITION 4.1. *For any trace operator T' of the form* (2.3) *there exist operators Λ and Λ^{-1} of the form* (4.4) *such that* (4.1) *can be replaced by* (4.3), *with $v=\Lambda u$, the operators described by* (4.5)-(4.6). □

However there is also a special variant adapted particularly to the Dirichlet problem that is more convenient for the stabilization problem.

Assume, for the moment, that 0 is not an eigenvalue of A_γ. (In the case where 0 is an eigenvalue, we replace A by $A+\delta$ for a small constant δ, carry out the constructions for this, and remove δ afterwards; see the explanation after Remark 5.3.) Let K_γ be the Poisson operator that solves the Dirichlet problem for A, i.e., K_γ maps φ into the solution u of

$$Au = 0 \quad \text{in } \Omega, \qquad \gamma u = \varphi \quad \text{on } \Gamma. \tag{4.7}$$

Since $K_\gamma T'$ has finite rank, the bounded operator $1-K_\gamma T'$ is a Fredholm operator with index 0 in $L^2(\Omega)$, and it maps $H^{2m}(\Omega)$ into $H^{2m}(\Omega)$.

Since

$$\gamma(1-K_\gamma T')u = \gamma u - T'u = Tu, \tag{4.8}$$

we have that

$$(1-K_\gamma T')D(A_T) \subseteq D(A_\gamma), \tag{4.9}$$

and moreover, if $u \in D(A_T)$ and $v = (1 - K_\gamma T')u$, then

$$Av = A(1 - K_\gamma T')u = Au. \tag{4.10}$$

Now we assume that T' can be chosen such that $1 - K_\gamma T'$ is a bijection in $H^{2m}(\Omega)$ (this will be done later; see § 5). Then $1 - K_\gamma T'$ maps $D(A_T)$ bijectively onto $D(A_\gamma)$, and hence defines a homeomorphism:

$$1 - K_\gamma T'\colon \quad D(A_T) \tilde{\to} D(A_\gamma) \qquad (= H^{2m}(\Omega) \cap H^m(\Omega)). \tag{4.11}$$

We have then established the useful factorization

$$A_T = A_\gamma(1 - K_\gamma T'), \tag{4.12}$$

in a precise sense. (See also Remark 1.1 in [8].)

Now proceeding as above, the problems

$$\partial_t^l u + A_T u = 0, \qquad u \in D(A_T), \quad l = 1, 2 \tag{4.13}$$

transform into

$$\partial_t^l v + (1 - K_\gamma T')A_\gamma v = 0, \qquad v \in D(A_\gamma), \quad l = 1, 2, \tag{4.14}$$

where $v = (1 - K_\gamma T')u$.

Thus we have transformed the perturbation of the second kind (4.13) into a perturbation of the first kind (4.14), and we are able to calculate the system operator in an easy way.

We have thus obtained Theorem 4.2.

THEOREM 4.2. *Assume that* $0 \notin \mathrm{sp}(A_\gamma)$. *The boundary feedback systems*

$$\begin{aligned} &\partial_t u + Au = 0 && \text{in } \Omega && \text{for } t > 0, \\ &\gamma u = T'u && \text{on } \Gamma && \text{for } t > 0, \\ &u = u_0 && \text{in } \Omega && \text{at } t = 0, \end{aligned} \tag{4.15}$$

and

$$\begin{aligned} &\partial_t^2 u + Au = 0 && \text{in } \Omega && \text{for } t \in \mathbb{R}, \\ &\gamma u = T'u && \text{on } \Gamma && \text{for } t \in \mathbb{R}, \\ &u = u_0 && \text{in } \Omega && \text{at } t = 0, \\ &\partial_t u = u_1 && \text{in } \Omega && \text{at } t = 0, \end{aligned} \tag{4.16}$$

with T' *given by* (2.3), *transform into the systems*

$$\begin{aligned} &\partial_t v + Av - K_\gamma T'Av = 0 && \text{in } \Omega && \text{for } t > 0, \\ &\gamma v = 0 && \text{on } \Gamma && \text{for } t > 0, \\ &v = v_0 && \text{in } \Omega && \text{for } t = 0, \end{aligned} \tag{4.15$'$}$$

and

$$\begin{aligned} &\partial_t^2 v + Av - K_\gamma T'Av = 0 && \text{in } \Omega && \text{for } t \in \mathbb{R}, \\ &\gamma v = 0 && \text{on } \Gamma && \text{for } t \in \mathbb{R}, \\ &v = v_0 && \text{in } \Omega && \text{for } t = 0, \\ &\partial_t v = v_1 && \text{in } \Omega && \text{for } t = 0. \end{aligned} \tag{4.16$'$}$$

Since

$$K_\gamma T'Av = \sum_{j=1}^{N} (Av \mid w_j) K_\gamma g_j, \tag{4.17}$$

for $v \in H^{2m}(\Omega)$, $K_\gamma T'A$ has finite rank, and we see that

$$\tilde{A} = A - K_\gamma T'A, \tag{4.18}$$

can be regarded as a finite-dimensional perturbation of A.

We obviously have ($\|\cdot\|_s$ is the $H^s(\Omega)$-norm):

$$\|K_\gamma T'Av\|_0 \leqq c\|Av\|_0 \leqq c\|Av\|_0 + \|v\|_0 \tag{4.19}$$

for $v \in D(A_\gamma)$, so $K_\gamma T'A$ is A-bounded. Since A_γ is the infinitesimal generator of an analytic semigroup on $L^2(\Omega)$, so is $\tilde{A}_\gamma$, from the perturbation result in Proposition 1 of Zabczyk [17].

We have thus obtained Theorem 4.3.

THEOREM 4.3. *The realization $\tilde{A}_\gamma$ of the operator*

$$\tilde{A} = A - K_\gamma T'A, \tag{4.20}$$

with domain

$$D(\tilde{A}_\gamma) = H^{2m}(\Omega) \cap H_0^m(\Omega) \quad (= D(A_\gamma)), \tag{4.21}$$

is the infinitesimal generator of an analytic semigroup $e^{-\tilde{A}_\gamma t}$, $t \geqq 0$ on $L^2(\Omega)$, giving the solution to (4.15′) *as*

$$v(t, x) = e^{-\tilde{A}_\gamma t} v_0(x), \qquad x \in \Omega, \quad t \geqq 0, \tag{4.22}$$

when $v_0 \in L^2(\Omega)$. The solution to the original system (4.15) *is then*

$$u(t, x) = (1 - K_\gamma T')^{-1} e^{-\tilde{A}_\gamma t}(1 - K_\gamma T')u_0(x), \qquad x \in \Omega, \quad t \geqq 0 \tag{4.23}$$

when $u_0 \in L^2(\Omega)$. A similar result holds of course for the realization $(A+G)_\gamma$ of the operator discussed in (4.3)–(4.6).

Remark 4.4. Both problems $(\partial_t + A_T)u = 0$ and $(\partial_t + \tilde{A}_\gamma)u = 0$ are special cases of the general parabolic pseudodifferential boundary value problems treated in [5]. It is shown in Theorem 4.1.1 of [5] that the solution operator is an analytic semigroup.

5. An application of the pseudodifferential transformation to stabilization. We will now show how the transformation from § 4 can be used as a shortcut to some of the results of Lasiecka and Triggiani [7]–[9], which have been a great motivation to us.

The assumed instability of the systems (1.1) and (1.2) is caused by the negative eigenvalues in the pure point spectrum $\operatorname{sp}(A_\gamma)$ of A_γ, and we will show that we can choose a finite-dimensional feedback boundary condition

$$\gamma u = T'u, \tag{5.1}$$

where T' is defined by

$$T'u = \sum_{j=1}^{N} (u \mid w_j) g_j \tag{5.2}$$

(see (2.4)), such that the systems

$$\begin{aligned} &\partial_t u + Au = 0 && \text{in } \Omega && \text{for } t > 0,\\ &\gamma u = T'u && \text{in } \Gamma && \text{for } t > 0,\\ &u = u_0 && \text{in } \Omega && \text{for } t = 0, \end{aligned} \tag{5.3}$$

and

$$(5.4)\quad \begin{aligned} &\partial_t^2 u + Au = 0 && \text{in } \Omega && \text{for } t \in \mathbb{R}, \\ &\gamma u = T'u && \text{on } \Gamma && \text{for } t \in \mathbb{R}, \\ &u = u_0 && \text{in } \Omega && \text{at } t = 0, \\ &\partial_t u = u_1 && \text{in } \Omega && \text{at } t = 0, \end{aligned}$$

are *stable* systems (in the sense described in § 1). We will apply the pseudodifferential transformation to the perturbations of the second kind (5.3) and (5.4) and then apply Wonham's "pole assignment theorem" (Wonham [16]) on the resulting perturbations of the first kind. This, combined with a resolvent analysis, gives us the desired results.

Let the eigenvalues of A_γ be arranged in a nondecreasing sequence

$$(5.5)\qquad \lambda_1 \leqq \lambda_2 \leqq \cdots \leqq \lambda_{K-1} \leqq 0 < \lambda_K \leqq \cdots,$$

each eigenvalue repeated according to multiplicity, and let $\{\varphi_j\}_{j\geqq 1}$ be a corresponding set of orthonormalized eigenfunctions of A_γ. Now define P_u and P_s as the orthogonal projections of $L^2(\Omega)$ on the orthogonal subspaces X_u, respectively X_s, defined by

$$(5.6)\qquad X_u = \operatorname{span}\{\varphi_j\}_{1\leqq j<K}, \qquad X_s = \overline{\operatorname{span}}\{\varphi_j\}_{j\geqq K}.$$

Remark 5.1. The results in Lasiecka and Triggiani [7]-[9] are formulated as if nonself-adjoint realizations are treated as well, but on the other hand, the treatment is based heavily on the orthogonal projections on the eigenspaces X_u and X_s. Orthogonality of eigenspaces in general requires at least that A_γ is normal, i.e., $A_\gamma A_\gamma^* = A_\gamma^* A_\gamma$, but we know of no Dirichlet realization A_γ satisfying this without being self-adjoint. □

Since X_u and $X_s \cap D(A_\gamma)$ are invariant subspaces for A_γ, we can define the restrictions

$$(5.7)\qquad A_u = A_\gamma|_{X_u}, \qquad A_s = A_\gamma|_{X_s \cap D(A_\gamma)}.$$

Then A_u is a bounded operator on X_u and A_s is an unbounded operator with domain $D(A_s) = X_s \cap D(A_\gamma)$. Note that P_u and P_s commute with A_γ on $D(A_\gamma)$. Now writing $f_u = P_u f$, $f_s = P_s f$ for $f \in L^2(\Omega)$, we have that when $u \in D(A_T)$ ((see 3.1) and (4.11)), then $v = (1 - K_\gamma T')u \in D(A_\gamma)$ satisfies

$$(5.8)\qquad Av = Au, \quad v_u \in X_u, \quad v_s \in D(A_s).$$

(Note that $v_s \in D(A_s)$ despite the fact that u_s does not necessarily belong to $D(A_s)$.)

Instead of working with the integral representation of the solution operator for the unperturbed problem as in Lasiecka and Triggiani [8], we attack the resolvents directly and in this way avoid negative fractional power domains.

We use the factorization

$$(5.9)\qquad A_T = A_\gamma(1 - K_\gamma T')$$

in the discussion of the resolvent equation

$$(5.10)\qquad (A_T - \lambda)u = f, \qquad f \in L^2(\Omega).$$

First we consider the case where we are allowed to *decouple* the feedback by assuming that

$$(5.11)\qquad P_s w_j = 0, \qquad j = 1, 2, \cdots, N$$

(i.e., the w_j are in X_u; the "unstable" eigenspace).

Then we write (5.10) in projected and factorized form

$$\begin{pmatrix} P_u \\ P_s \end{pmatrix}(A_\gamma(1-K_\gamma T')(u_u+u_s)-\lambda(u_u+u_s))=\begin{pmatrix} f_u \\ f_s \end{pmatrix} \tag{5.12}$$

and we compute, for $u \in D(A_T)$:

$$\begin{aligned} P_uA_\gamma(1-K_\gamma T')(u_u+u_s)-P_u\lambda(u_u+u_s) &= A_\gamma P_u(1-K_\gamma T')(u_u+u_s)-\lambda u_u \\ &= A_u u_u - A_u P_u K_\gamma T' u_u - \lambda_{u_u}, \\ P_sA_\gamma(1-K_\gamma T')(u_u+u_s)-P_s\lambda(u_u+u_s) &= A_\gamma P_s(1-K_\gamma T')(u_u+u_s)-\lambda u_s \\ &= A_s(u_s - P_s K_\gamma T' u_u) - \lambda u_s. \end{aligned}$$

Since $u_s - P_s K_\gamma T' u_u = v_s$ belongs to $D(A_s)$, the factorization (5.12) is legimate, and (5.12) reduces to

$$A_u u_u - A_u P_u K_\gamma T' u_u - \lambda u_u = f_u, \tag{5.13}$$

$$A_s(u_s - P_s K_\gamma T' u_u) - \lambda u_s = f_s \tag{5.14}$$

(see also (5.42)), where we observe that (5.13) is a finite-dimensional resolvent equation for the matrix operator

$$\bar{A}_u = A_u - A_u P_u K_\gamma T'. \tag{5.15}$$

At this point we can use the same arguments as Lasiecka and Triggiani [8] to get a good choice of T'. We give the full details in the most straightforward case, and refer to Lasiecka and Triggiani for partial information on other cases.

PROPOSITION 5.2. *Assume that the Neumann traces* $\{\nu\varphi_j\}_{1\leqq j<K}$ *are linearly independent, so that*

$$\dim(\nu X_u) = \dim(X_u) \quad (=K-1), \tag{5.16}$$

and let $\{c_j\}_{1\leqq j<K}$ *be an arbitrarily given set of* $K-1$ *distinct, real numbers.*

Then there exists a number N *and a set*

$$\{w_j, g_j\}_{1\leqq j \leqq N}, \tag{5.17}$$

where $w_j \in X_u$ *and* $g_j \in C^\infty(\Gamma)^m$, *such that with*

$$T'u = \sum_{j=1}^{N} (u \mid w_j) g_j, \tag{5.18}$$

the eigenvalues of the matrix operator $\bar{A}_u$, *defined by*

$$\bar{A}_u v = (A_u - A_u P_u K_\gamma T')v \quad \textit{for } v \in X_u, \tag{5.19}$$

are precisely the set $\{c_j\}_{1\leqq j<K}$.

The number N *can be taken as the largest multiplicity of the unstable eigenvalues* $\{\lambda_j\}_{1\leqq j<K}$. *In particular,* $N=1$ *when the eigenvalues are simple.*

Proof. Assume first that *all* of the eigenvalues $\{\lambda_j\}_{1\leqq j<K}$ are simple and take $N=1$. Consider T' of the form

$$T'u = (u \mid w) g. \tag{5.20}$$

In the basis $\{\varphi_j\}_{1\leqq j<K}$ of X_u, the matrix $\bar{A}_u$ has the form

$$\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \lambda_{K-1} \end{pmatrix} - P(g)W^t, \tag{5.21}$$

where $P(g)$ is the column vector

$$(5.22)\qquad P(g)=\begin{pmatrix} \lambda_1(K_\gamma g\,|\,\varphi_1) \\ \lambda_2(K_\gamma g\,|\,\varphi_2) \\ \vdots \\ \lambda_{K-1}(K_\gamma g\,|\,\varphi_{K-1}) \end{pmatrix}$$

and W^t is the transpose of the column vector W given by

$$(5.23)\qquad W=\begin{pmatrix} (\varphi_1\,|\,w) \\ (\varphi_2\,|\,w) \\ \vdots \\ (\varphi_{K-1}\,|\,w) \end{pmatrix}.$$

Consider now the control matrix of the pair (A_u, W):

$$[W \quad A_u W \quad \cdots \quad A_u^{K-2}W]$$

$$(5.24)\qquad =\begin{pmatrix} (\varphi_1\,|\,w) & \lambda_1(\varphi_1\,|\,w) & \cdots & \lambda_1^{K-2}(\varphi_1\,|\,w) \\ (\varphi_2\,|\,w) & \lambda_2(\varphi_2\,|\,w) & \cdots & \lambda_2^{K-2}(\varphi_2\,|\,w) \\ \vdots & \vdots & & \vdots \\ (\varphi_{K-1}\,|\,w) & \lambda_{K-1}(\varphi_{K-1}\,|\,w) & \cdots & \lambda_{K-1}^{K-2}(\varphi_{K-1}\,|\,w) \end{pmatrix}.$$

The determinant of the control matrix is calculated to be

$$(5.25)\qquad \prod_{1\leqq j<K}(\varphi_j\,|\,w)\prod_{1\leqq l<k<K}(\lambda_k-\lambda_l),$$

by reduction to a Vandermonde determinant. Since the eigenvalues are assumed to be simple, we can choose $w\in X_u$, satisfying $(\varphi_j\,|\,w)\neq 0$ for $j=1,2,\cdots,K-1$, such that the determinant is different from 0. This implies that the pair (A_u, W) is controllable, so by the pole assignment theorem (Wonham [16]) there exists a matrix

$$(5.26)\qquad \tilde{P}=\begin{pmatrix} p_1 \\ p_2 \\ \vdots \\ p_{K-1} \end{pmatrix}, \qquad p_j\in\mathbb{C}, \quad j=1,2,\cdots,K-1,$$

for which the matrix $A_u-\tilde{P}W^t$ has the set $\{c_j\}_{1\leqq j<K}$ as eigenvalues. Now we will choose $g\in C^\infty(\Gamma)^m$ such that $\tilde{P}=P(g)$, i.e., such that

$$(5.27)\qquad \lambda_j(K_\gamma g\,|\,\varphi_j)=p_j,$$

for $j=1,2,\cdots,K-1$.

From the formula (A10) in the Appendix we have that

$$(5.28)\qquad (K_\gamma g\,|\,\varphi_j)=\frac{-1}{\lambda_j}(g\,|\,\mathscr{A}^{10*}\nu\varphi_j)_\Gamma.$$

Here $\mathscr{A}^{10}$ is (since A is elliptic) an invertible $m\times m$ matrix of differential operators over Γ, that maps $\prod_{m\leqq k<2m}H^{2m-k-1/2}(\Gamma)$ onto $\prod_{0\leqq k<m}H^{m+k+1/2}(\Gamma)$, and $(\cdot\,|\,\cdot)_\Gamma$ is the $L^2(\Gamma)^m$ inner product.

Since the set

$$(5.29)\qquad \{\nu\varphi_1, \nu\varphi_2, \cdots, \nu\varphi_{K-1}\},$$

is linearly independent, so also is the set

$$\{\mathscr{A}^{10^*}\nu\varphi_1, \mathscr{A}^{10^*}\nu\varphi_2, \cdots, \mathscr{A}^{10^*}\nu\varphi_{K-1}\}. \tag{5.30}$$

Hence it is possible to choose $g \in C^\infty(\Gamma)^m$, satisfying

$$(g \mid \mathscr{A}^{10*}\nu\varphi_j) = -p_j, \qquad j = 1, 2, \cdots, K-1, \tag{5.31}$$

and this choice of g gives us the components of the desired $P(g)$, in view of (5.28). This ends the proof in the case of simple eigenvalues.

Now assume that one or more of the eigenvalues $\{\lambda_j\}_{1\leqq j<K}$ have multiplicity larger than 1, and let us take σ to be the largest occurring multiplicity. Take $N=\sigma$ and consider T' of the form

$$T'u = \sum_{j=1}^{\sigma} (u \mid w_j) g_j. \tag{5.32}$$

In this case $A_u - A_u P_u K_\gamma T'$ can be written

$$\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & & \cdots & \lambda_{K-1} \end{pmatrix} - P(\{g_i\}_{1\leqq i\leqq\sigma})\, W_\sigma^t, \tag{5.33}$$

where $P(\{g_i\}_{1\leqq i\leqq\sigma})$ is the $(K-1)\times\sigma$ matrix

$$P(\{g_i\}_{1\leqq i\leqq\sigma}) = \begin{pmatrix} \lambda_1(K_\gamma g_1 \mid \varphi_1) & \lambda_1(K_\gamma g_2 \mid \varphi_2) & \cdots & \lambda_1(K_\gamma g_\sigma \mid \varphi_1) \\ \lambda_2(K_\gamma g_1 \mid \varphi_2) & \lambda_2(K_\gamma g_2 \mid \varphi_2) & \cdots & \lambda_2(K_\gamma g_\sigma \mid \varphi_2) \\ \vdots & \vdots & & \vdots \\ \lambda_{K-1}(K_\gamma g_1 \mid \varphi_{K-1}) & \lambda_{K-1}(K_\gamma g_2 \mid \varphi_{K-1}) & \cdots & \lambda_{K-1}(K_\gamma g_\sigma \mid \varphi_{K-1}) \end{pmatrix} \tag{5.34}$$

and W_σ is the $(K-1)\times\sigma$ matrix

$$W_\sigma = \begin{pmatrix} (\varphi_1 \mid w_1) & (\varphi_1 \mid w_2) & \cdots & (\varphi_1 \mid w_\sigma) \\ (\varphi_2 \mid w_1) & (\varphi_2 \mid w_2) & \cdots & (\varphi_2 \mid w_\sigma) \\ \vdots & & & \\ (\varphi_{K-1} \mid w_1) & (\varphi_{K-1} \mid w_2) & \cdots & (\varphi_{K-1} \mid w_\sigma) \end{pmatrix}. \tag{5.35}$$

Considering the form of the control matrix

$$[W_\sigma \quad A_u W_\sigma \quad \cdots \quad A_u^{K-2} W_\sigma], \tag{5.36}$$

we see that if $w_1, w_2, \cdots, w_\sigma$ are chosen in X_u such that

$$\operatorname{rank} W_\sigma = \sigma, \tag{5.37}$$

then the rank of the control matrix (5.36) is $K-1$ (because a regular $(K-1)\times(K-1)$ submatrix can be extracted, after suitable row-column operations). Then according to Wonham's theorem, there exists a complex matrix

$$\tilde{P}_\sigma = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1\sigma} \\ p_{21} & p_{22} & \cdots & p_{2\sigma} \\ \vdots & \vdots & & \vdots \\ p_{K-1,1} & p_{K-1,2} & \cdots & p_{K-1,\sigma} \end{pmatrix} \tag{5.38}$$

such that the eigenvalues of the matrix $A_u - \tilde{P}_\sigma W_\sigma^t$ are $\{c_j\}_{1\leqq j<K}$.

To obtain $\tilde{P}_\sigma = P(\{g_i\}_{1 \leqq i \leqq \sigma})$ we use that, as in (5.28),

$$(K_\gamma g_i \mid \varphi_j) = \frac{-1}{\lambda_j}(g_i \mid \mathscr{A}^{10^*}\nu\varphi_j)_\Gamma. \tag{5.39}$$

In view of (5.16) we can choose $g_i \in C^\infty(\Gamma)^m$, $i = 1, 2, \cdots, \sigma$, satisfying

$$(g_i \mid \mathscr{A}^{10*}\nu\varphi_j)_\Gamma = -p_{ji}, \qquad j = 1, 2, \cdots, K-1, \quad i = 1, 2, \cdots, \sigma, \tag{5.40}$$

and this choice of $\{g_i\}_{1 \leqq i \leqq \sigma}$ provides us with the desired $P(\{g_i\}_{1 \leqq i \leqq \sigma})$. □

Remark 5.3. For the application of the Wonham theorem here, it is important that the range of $P_u K_\gamma$ fills out all of X_u; this is reformulated as the question of whether the Neumann traces of the Dirichlet eigenfunctions in X_u are linearly independent. In that case the results are easy to formulate and allow N to be very low; otherwise, the results become increasingly complicated and require (in general) higher N, the more linear dependence there is. Lasiecka and Triggiani have in [8] an upper and lower bound on the number N of feedback terms necessary in (5.18), once the number of linearly independent Neumann traces are given, but the discussion of the size of $\dim(\nu X_u)$ for differential operators on general domains in $\mathbb{R}^n$ in the literature is far from complete, as far as we know.

Let us now dispose of the temporary assumption $0 \notin \operatorname{sp}(A_\gamma)$ from § 4. Everything done thus far can be done without changes with the translated operator $A_\gamma - \delta$, $\delta > 0$, so that if $0 \in \operatorname{sp}(A_\gamma)$, we take a $\delta > 0$ such that $0 \notin \operatorname{sp}(A_\gamma - \delta)$. Then after determining the new realization $A_T - \delta$ that moves the eigenvalues $\lambda_1 - \delta, \lambda_2 - \delta, \cdots, \lambda_{K-1} - \delta$, and leaves $\lambda_j - \delta, j \geqq K$ unaffected, we just add δ to $A_T - \delta$. Then the moved eigenvalues will be increased with δ, and so will the unaffected eigenvalues; i.e., they will return to be the eigenvalues of A_s.

Let us now choose the set $\{c_j\}_{1 \leqq j < K}$ occurring in Proposition 5.2 such that $c_j \geqq \lambda_K (>0), j = 1, 2, \cdots, K-1$. With T' chosen according to the theorem, the operator $P_u(1 - K_\gamma T')$ is injective, hence bijective, on X_u, and since $w_j \in X_u$, $1 - K_\gamma T'$ is the identity on X_s. Then, as promised in § 4, $1 - K_\gamma T'$ is bijective from $H^{2m}(\Omega)$ to $H^{2m}(\Omega)$ ($K_\gamma T'$ has C^∞-range) and maps $D(A_T)$ onto $D(A_\gamma)$. This justifies the factorization (4.12). Define $R(\lambda, \bar{A}_u)$ as the resolvent of $\bar{A}_u$ in X_u, for all $\lambda \notin \{c_j\}_{1 \leqq j < K}$ and let $R(\lambda, A_s)$ be the resolvent of A_s in X_s, defined for all $\lambda \notin \{\lambda_j\}_{j \geqq K}$.

We can then write the solution to (5.13) as

$$u_u = R(\lambda, \bar{A}_u) f_u, \qquad \lambda \notin \{c_j\}_{1 \leqq j < K}. \tag{5.41}$$

We see that if $u \in D(A_T)$, then $v = (1 - K_\gamma T')u$ belongs to $D(A_\gamma)$, hence $v_s \in D(A_s)$. Since

$$\begin{aligned} v_s &= P_s(1 - K_\gamma T')u = P_s(1 - K_\gamma T')(u_u + u_s) \\ &= u_s - P_s K_\gamma T' u_u, \end{aligned} \tag{5.42}$$

we see that $u_s - P_s K_\gamma T' u_u \in D(A_s)$, so (5.14) is justified and can be written as

$$(A_s - \lambda)(u_s - P_s K_\gamma T' u_u) = f_s + \lambda P_s K_\gamma T' u_u. \tag{5.43}$$

For all $\lambda \notin \{\lambda_j\}_{j \geqq K}$ we have

$$u_s - P_s K_\gamma T' u_u = R(\lambda, A_s)(f_s + \lambda P_s K_\gamma T' u_u). \tag{5.44}$$

Inserting (5.41), we find for all $\lambda \notin (\{c_j\}_{1 \leqq j < K} \cup \{\lambda_j\}_{j \geqq K})$:

$$u_s = P_s K_\gamma T' R(\lambda, \bar{A}_u) f_u + R(\lambda, A_s)(f_s + \lambda P_s K_\gamma T' R(\lambda, \bar{A}_u) f_u), \tag{5.45}$$

so that

$$
\begin{aligned}
u &= u_u + u_s \\
&= R(\lambda, \bar{A}_u)f_u + P_sK_\gamma T'R(\lambda, \bar{A}_u)f_u + R(\lambda, A_s)f_s \\
&\quad + \lambda R(\lambda, A_s)P_sK_\gamma T'R(\lambda, \bar{A}_u)f_u \\
&= (1 + P_sK_\gamma T' + \lambda R(\lambda, A_s)P_sK_\gamma T')R(\lambda, \bar{A}_u)f_u + R(\lambda, A_s)f_s.
\end{aligned}
\tag{5.46}
$$

We have now obtained Proposition 5.4.

PROPOSITION 5.4. *The resolvent* $R(\lambda, A_T)$ *solving* (5.13)-(5.14) *can be written in the form*

$$R(\lambda, A_T)f = (R_{11} \quad R_{12})\begin{pmatrix} f_u \\ f_s \end{pmatrix}. \tag{5.47}$$

where

$$
\begin{aligned}
R_{11} &= (1 + P_sK_\gamma T' + \lambda R(\lambda, A_s)P_sK_\gamma T')R(\lambda, \bar{A}_u), \\
R_{12} &= R(\lambda, A_s),
\end{aligned}
\tag{5.48}
$$

so $R(\lambda, A_T)$ *is well defined for all* λ *outside the spectrum of* A_T, *and maps* $L^2(\Omega)$ *into* $H^{2m}(\Omega)$.

Let us then prove Proposition 5.5.

PROPOSITION 5.5. *Assume that the prescribed eigenvalues* $\{c_j\}_{1 \leqq j < K}$ *of* $\bar{A}_u$ *are chosen such that* $c_j \geqq \lambda_K$, $j = 1, \cdots, K-1$. *There is then a constant* $M_T > 0$, *independent of* λ, *such that the resolvent* $R(\lambda, A_T)$ *satisfies the inequality*

$$\|R(\lambda, A_T)\|_{L^2, L^2} \leqq \frac{M_T}{\operatorname{dist}(\lambda, I_K)} \tag{5.49}$$

as an operator in $L^2(\Omega)$. *Here* $I_K = [\lambda_K, \infty[$, *which in this case equals the closed convex hull of the spectrum of* A_T.

Proof. A_s is a self-adjoint, positive operator on $X_s \cap D(A_\gamma)$, satisfying

$$
\begin{aligned}
\|(A_s - \lambda)u\|_0\|u\|_0 &\geqq |((A_s - \lambda)u \,|\, u)| \\
&= |((A_s - \operatorname{Re}\lambda)u \,|\, u) - i \operatorname{Im}\lambda \|u\|_0^2| \\
&= (((A_s - \operatorname{Re}\lambda)u \,|\, u)^2 + (\operatorname{Im}\lambda \|u\|_0)^2)^{1/2} \\
&\geqq \begin{cases} |\operatorname{Im}\lambda| \|u\|_0^2 & \text{if } \operatorname{Re}\lambda \geqq \lambda_K \\ ((\lambda_K - \operatorname{Re}\lambda)^2 + (\operatorname{Im}\lambda)^2)^{1/2}\|u\|_0^2 & \text{if } \operatorname{Re}\lambda \leqq \lambda_K \end{cases} \\
&\geqq \operatorname{dist}(\lambda, I_K)\|u\|_0^2;
\end{aligned}
$$

hence,

$$\|R(\lambda, A_s)\|_{L^2, L^2} \leqq \operatorname{dist}(\lambda, I_K)^{-1}. \tag{5.50}$$

$R(\lambda, \bar{A}_u)$ is a $(K-1) \times (K-1)$ matrix of the form

$$R(\lambda, \bar{A}_u) = \det(\bar{A}_u - \lambda)^{-1} p(\lambda, \bar{A}_u), \tag{5.51}$$

where $p(\lambda, \bar{A}_u)$ is a polynomial in λ (this follows easily from the inversion formula for matrices), and

$$\det(\bar{A}_u - \lambda) = \text{const.} \cdot \prod_{1 \leqq j < K} (\lambda - c_j). \tag{5.52}$$

Therefore $\|R(\lambda, \bar{A}_u)\|_{L^2,L^2}$ is $O(|\lambda - c_k|^{-1})$ in a neighbourhood of each $c_k \in \{c_j\}_{1\leqq j<K}$. For $|\lambda\gamma \to \infty$ we write

$$\bar{A}_u - \lambda = -\lambda(1 - \lambda^{-1}\bar{A}_u) \tag{5.53}$$

and we see that $R(\lambda, \bar{A}_u)$ is $O(|\lambda|^{-1})$ for $|\lambda| \to \infty$, since $(1-\lambda^{-1}\bar{A}_u)^{-1}$ can be expressed by a Neumann series in $\lambda^{-1}\bar{A}_u$ for $|\lambda| > \|\bar{A}_u\|_{L^2,L^2}$. We can then conclude that

$$\|R(\lambda, \bar{A}_u)\|_{L^2,L^2} \leqq M(\text{dist}\,(\lambda, \text{sp}\,(\bar{A}_u)))^{-1}, \tag{5.54}$$

where M is a constant independent of λ.

Obviously, $\text{sp}\,(\bar{A}_u) \subset I_K$, and since $\text{sp}\,(\bar{A}_u)$ is bounded, $\|\lambda R(\lambda \bar{A}_u)\|_{L^2,L^2}$ is $O(1)$ for $|\lambda| \to \infty$; moreover, $P_s K_\gamma T'$ is a bounded operator, and the decomposition $f \to f_u + f_s$ is bounded and λ-independent. Then, from the form of $R(\lambda, A_T)$ (5.47)-(5.48), we find that

$$\begin{aligned} \|R(\lambda, A_T)\|^2_{L^2,L^2} &\leqq \frac{M'^2}{\text{dist}\,(\lambda, \text{sp}\,(\bar{A}_u))^2} + \frac{1}{\text{dist}\,(\lambda, I_K)^2} \\ &\leqq \frac{M_T^2}{\text{dist}\,(\lambda, I_K)^2}, \end{aligned} \tag{5.55}$$

where M' and M_T are positive constants. □

Using Lemma 3.1, Proposition 5.2, and Proposition 5.5 we find Theorem 5.6.

THEOREM 5.6. *Assume that*

(1) *The Neumann traces* $\{\nu\varphi_j\}_{1\leqq j<K}$ *of the Dirichlet eigenfunctions* φ_j *are linearly independent* (*but see also Remark* 5.3);

(2) *The largest occurring multiplicity of the Dirichlet eigenvalues* $\{\lambda_j\}_{1\leqq j<K}$ *of* A *is* N. *Then there exists a finite-dimensional boundary condition*

$$\gamma u = T'u \quad on\ \Gamma \tag{5.56}$$

where

$$T'u = \sum_{j=1}^{N} (u \,|\, w_j) g_j. \tag{5.57}$$

$w_j \in X_u$, $g_j \in C^\infty(\Gamma)^m$, $j = 1, 2, \cdots, N$, *such that the realization* A_T *of* A, *with domain*

$$D(A_T) = \{u \in H^{2m}(\Omega) \,|\, Tu = \gamma u - T'u = 0\}, \tag{5.58}$$

is the infinitesimal generator of an analytic semigroup $e^{-A_T t}$, $t \geqq 0$ *on* $L^2(\Omega)$, *giving the solution to the Dirichlet boundary feedback parabolic system* (5.3) *as*

$$u(t, x) = e^{-A_T T} u_0(x), \qquad x \in \Omega, \quad t \geqq 0, \tag{5.59}$$

when $u_0 \in L^2(\Omega)$, *and such that the solution* (5.59) *satisfies the damping estimate*

$$\|u(t, \cdot)\|_0 \leqq M e^{-\lambda_K t} \|u_0\|_0, \qquad t \geqq 0, \quad M > 0, \tag{5.60}$$

where λ_K *is the first positive Dirichlet eigenvalue of* A. *Moreover, the operators*

$$\cos\,(A_T^{1/2} t) \quad and \quad \sin\,(A_T^{1/2} t), \tag{5.61}$$

are well defined, and we can write the solution to the hyperbolic problem (5.4) *as*

$$u(t, x) = \cos\,(A_T^{1/2} t) u_0 + A_T^{-1/2} \sin\,(A_T^{1/2} t) u_1(x), \tag{5.62}$$

$x \in \Omega$, $t \in \mathbb{R}$, *when* $u_0, u_1 \in L^2(\Omega)$.

Remark 7. We see that in the "decoupled" case where $w_j \in X_u$, $j = 1, \cdots, K-1$, the estimate (5.60) holds. When $P_s w_j \neq 0$, the damping coefficient λ_K must be substituted by $\lambda_K - \varepsilon$, as we shall see later. Lasiecka and Triggiani did not mention the sharper estimate (5.60) for the decoupled case. Moreover, one of the slightly mysterious facts about the perturbation of the second kind is that the operator A_T can never be self-adjoint. This is a consequence of Proposition 1.7.11 in Grubb [5]. What is shown there is, more precisely, that A_T cannot even be weakly semibounded, i.e., cannot even satisfy

$$\text{(5.63)} \qquad \operatorname{Re} e^{i\theta}(A_T u \mid u) \geqq -c\|u\|_m^2 \quad \text{for } u \in D(A_T),$$

some c and θ, when $T' \neq 0$. It may be of interest to observe that Neumann problems behave differently as follows.

If A_T is the realization defined by a Neumann feedback boundary condition $\nu u + S\gamma u = T'u$, then (5.63) holds for general S and T' (for the conditions in Theorem 1.7.13 of [5] are void in the Neumann case). Let us also recall from [5, § 1.7] that weak semiboundedness (5.63) assures m-boundedness

$$\text{(5.64)} \qquad |(A_T u \mid v)| \leqq C\|u\|_m \|v\|_m \quad \text{for } u \in D(A_T),$$

for normal pseudodifferential boundary problems.

Since in the Dirichlet case, the realization A_T is never semibounded, the semigroup $e^{-A_T t}$, $t \geqq 0$ is never a *contraction* semigroup, hence the constant M in (5.60) is always greater than 1. This has also been noted by Lasiecka and Triggiani, who considered the translated Laplacian in [8].

Remark 5.8. Comparison of (5.59) and (4.23) show that for the semigroups we have

$$\text{(5.65)} \qquad e^{-A_T t} = (1 - K_\gamma T')^{-1} e^{-\tilde{A}_\gamma t}(1 - K_\gamma T').$$

This justifies the term "generalized change of coordinates" from § 2.

Now it is straightforward to extend the theory to include more general cases where $P_s w_j \neq 0$. The operator T' considered above can be written

$$\text{(5.66)} \qquad T'u = \sum_{j=1}^{N} (u \mid P_u w_j) g_j,$$

so if we let the w_j be arbitrary and define the operator T'' as

$$\text{(5.67)} \qquad T''u = \sum_{j=1}^{N} (u \mid P_s w_j) g_j,$$

we see that the decoupled case considered above corresponds to the case where $T'' = 0$.

The operator T_1, defined by

$$\text{(5.68)} \qquad T_1 u = \gamma u - T'u - T''u,$$

defines a normal boundary condition $T_1 u = 0$ (in the sense of Grubb [5]), just as T did, hence the operator realization A_{T_1} of A, with domain

$$\text{(5.69)} \qquad D(A_{T_1}) = \{u \in H^{2m}(\Omega) \mid T_1 u = 0\},$$

is a densely defined, closed operator in $L^2(\Omega)$. Here T_1 can be regarded as a perturbation of the trace operator $T = \gamma - T'$.

Let K_T be the Poisson solution operator defined by $u = K_T \varphi$, where u is the solution of

$$\text{(5.70)} \qquad Au = 0 \quad \text{in } \Omega, \qquad Tu = \varphi \quad \text{on } \Gamma.$$

We assume in the following that T (i.e., the sets $\{P_u w_j\}_{1 \leqq j < K}$ and $\{g_j\}_{1 \leqq j < K}$) is chosen such that the conclusions of Theorem 5.6 are valid, and then we study A_{T_1}.

Here we use that A_T is chosen to be bijective from $D(A_T)$ to $L^2(\Omega)$, which implies that K_T is well defined. Observe also the estimate

$$\begin{aligned}\|K_TT''u\|_0 &= \left\| \sum_{j=1}^{N} (u\,|\,P_sw_j)K_Tg_j \right\|_0 \\ &\leqq \|u\|_0 \sum_{j=1}^{N} \|P_sw_j\|_0\|K_Tg_j\|_0,\end{aligned} \tag{5.71}$$

where $\|K_Tg_j\|_0$ depends only on the sets $\{P_uw_j\}_{1\leqq j<K}$ and $\{g_j\}_{1\leqq j<J}$.

LEMMA 5.9. *Assume that the sets* $\{P_uw_j\}_{1\leqq j<K}$ *and* $\{g_j\}_{1\leqq j<K}$ *are chosen such that the conclusions of Theorem* 5.6 *are valid. Then there exists a constant* $r_1>0$, *such that for* $\|P_sw_j\|_0<r_1$, $j=1, 2, \cdots, N$ *the operator*

$$1-K_TT'' \tag{5.72}$$

is a homeomorphism in $L^2(\Omega)$ *and in* $H^{2m}(\Omega)$, *and, in particular, defines a bijection*

$$1-K_TT''\colon \quad D(A_{T_1}) \overset{\sim}{\to} D(A_T). \tag{5.73}$$

Moreover, when $u\in D(A_{T_1})$ *and* $v=(1-K_TT'')u$, *then* $Au=Av$, *in fact we have the factorization*

$$A_{T_1}=A_T(1-K_TT''). \tag{5.74}$$

Proof. Let $r_1>0$ be chosen such that for $\|P_sw_j\|_0<r_1$, $j=1, 2, \cdots, N$, we have $\|K_TT''\|_{L^2,L^2}\leqq\frac{1}{2}$. This is possible by (5.71). Now $1-K_TT''$ is a bounded operator in $L^2(\Omega)$ and is inverted by a Neumann series

$$(1-K_TT'')^{-1}=\sum_{m=0}^{\infty} (K_TT'')^m, \tag{5.75}$$

converging in the operator norm in $L^2(\Omega)$. Thus $1-K_TT''$ is a homeomorphism of $L^2(\Omega)$ onto itself.

Since K_T has range in $H^{2m}(\Omega)$, we see that $1-K_TT''$ is likewise a homeomorphism of $H^{2m}(\Omega)$, onto itself, and, since

$$T(1-K_TT'')u=Tu-T''u=T_1u, \tag{5.76}$$

$u\in D(A_{T_1})$ if and only if $v=(1-K_TT'')u\in D(A_T)$, so $1-K_TT''$ defines a bijection of $D(A_{T_1})$ onto $D(A_T)$. The last observation follows from the fact that $AK_T=0$. □

We will now study the resolvent $R(\lambda, A_{T_1})$ of A_{T_1}, and we start out with the equation

$$\begin{aligned}&(A-\lambda)u=f \quad \text{in } \Omega,\\ &T_1u=0 \quad \text{on } \Gamma.\end{aligned} \tag{5.77}$$

Using (5.73) with $v=(1-K_TT'')u$ we get

$$\begin{aligned}&(A-\lambda)(v+K_TT''u)=f \quad \text{in } \Omega,\\ &Tv=0 \quad \text{on } \Gamma,\end{aligned} \tag{5.78}$$

so that

$$\begin{aligned}&(A-\lambda)v=f+\lambda K_TT''u \quad \text{in } \Omega,\\ &Tv=0 \quad \text{on } \Gamma,\end{aligned} \tag{5.79}$$

i.e., if λ is in the resolvent set of A_T,

$$\begin{aligned} v &= R(\lambda, A_T)(f+\lambda K_T T''u) \quad \text{in } \Omega, \\ Tv &= 0 \quad \text{on } \Gamma. \end{aligned} \tag{5.80}$$

For u this gives

$$u - K_T T''u = v = R(\lambda, A_T)(f+\lambda K_T T''u), \tag{5.81}$$

so that

$$(1-(1+\lambda R(\lambda, A_T))K_T T'')u = R(\lambda, A_T)f. \tag{5.82}$$

Let us denote, for $\varepsilon > 0$, $0 < \theta < \pi/2$ the obtuse sector (disjoint from I_K)

$$W_{\lambda_K,\varepsilon,\theta} = \left\{ z \in \mathbb{C} \,\middle|\, z = (\lambda_K - \varepsilon) + r\, e^{i\omega},\ r \geqq 0, \frac{\pi}{2} - \theta < \omega < \frac{3\pi}{2} + \theta \right\}.$$

From the estimate (5.49) we see that $\lambda R(\lambda, A_T)$ is bounded on $W_{\lambda_K,\varepsilon,\theta}$, and hence for any $\varepsilon > 0$, any $\theta \in]0, \pi/2[$, there exists by (5.71) a constant $r > 0$ such that for $w_j \in L^2(\Omega)$, satisfying $\|P_s w_j\|_0 < r \leqq r_1$, $j = 1, 2, \cdots, N$, we have

$$\|(1+\lambda R(\lambda, A_T))K_T T''\|_{L^2,L^2} \leqq \tfrac{1}{2} \tag{5.83}$$

for all $\lambda \in W_{\lambda_K,\varepsilon,\theta}$.

With the w_j, $j = 1, 2, \cdots, N$, chosen in this way, the resolvent $R(\lambda, A_{T_1})$ of A_{T_1} is a well-defined, bounded operator in $L^2(\Omega)$ for $\lambda \in W_{\lambda_K,\varepsilon,\theta}$, given by (see (5.73))

$$\begin{aligned} R(\lambda, A_{T_1}) &= (1-(1+\lambda R(\lambda, A_T))K_T T'')^{-1} R(\lambda, A_T) \\ &= \sum_{m=0}^{\infty} ((1+\lambda R(\lambda, A_T))K_T T'')^m R(\lambda, A_T), \end{aligned} \tag{5.84}$$

and satisfying the estimate

$$\|R(\lambda, A_{T_1})\|_{L^2,L^2} \leqq \frac{c_1}{|\lambda - \lambda_K|}. \tag{5.85}$$

Here $c_1 > 0$ is a constant, independent of λ.

Altogether, we have obtained Theorem 5.10.

THEOREM 5.10. *Let $\varepsilon > 0$ be given, and assume that the sets $\{P_u w_j\}_{1 \leqq j < K}$ and $\{g_j\}_{1 \leqq j < K}$ are chosen such that the conclusions of Theorem 5.6 are valid. The finite-dimensional Dirichlet boundary feedback*

$$\gamma u = T'u + T''u = \sum_{j=1}^{N} (u \mid w_j) g_j \tag{5.86}$$

defines a realization A_{T_1} of A, with domain

$$D(A_{T_1}) = \{u \in H^{2m}(\Omega) \mid T_1 u = 0\}, \tag{5.87}$$

where T_1 is the trace operator defined by

$$T_1 = \gamma - T' - T'', \tag{5.88}$$

and there exists a constant $r>0$, *such that for arbitrary choices of* $P_s w_j$, *with* $\|P_s w_j\|_0 < r$, A_{T_1} *is the infinitesimal generator of an analytic semigroup* $e^{-A_{T_1}t}$, $t \geqq 0$, *on* $L^2(\Omega)$, *giving the solution to the Dirichlet boundary feedback control system*

$$(5.89)\qquad \begin{aligned} &\partial_t u + Au = 0 && \textit{in } \Omega \quad \textit{for } t>0, \\ &\gamma u = \sum_{j=1}^{N} (u\,|\,w_j) g_j && \textit{on } \Gamma \quad \textit{for } t>0, \\ &u = u_0 && \textit{in } \Omega \quad \textit{at } t=0, \end{aligned}$$

as

$$(5.90)\qquad u(t,x) = e^{-A_{T_1}t} u_0(x), \qquad t \geqq 0, \quad x \in \Omega, \quad u_0 \in L^2(\Omega),$$

where the solution (5.90) *satisfies*

$$(5.91)\qquad \|u(t,\cdot)\|_0 \leqq M' e^{-(\lambda_K - \varepsilon)t} \|u_0\|_0, \qquad t \geqq 0.$$

Here λ_K *is the first positive Dirichlet eigenvalue of* A, *and* M' *is a constant greater than zero.*

We will now use the factorization (5.74) in the investigation of the hyperbolic problem for A_{T_1}.

The boundary value problem

$$(5.92)\qquad \partial_t^2 u + A_{T_1} u = 0, \qquad u \in D(A_{T_1})$$

transforms by (5.74) into

$$(5.93)\qquad (1 - K_T T'')^{-1} \partial_t^2 v + A_T v = 0, \qquad v \in D(A_T)$$

where $v = (1 - K_T T'')u$.

Acting with $(1 - K_T T'')$ from the left in (5.93) we find

$$(5.94)\qquad \partial_t^2 v + (1 - K_T T'') A_T v = 0, \qquad v \in D(A_T).$$

Moreover, if we now impose on the w_j to satisfy $P_s w_j \in D(A_T^*)$, $j = 1, 2, \cdots, N$, then for $v \in D(A_T)$:

$$\begin{aligned} \|K_T T'' A_T v\|_0 &= \left\| \sum_{j=1}^{N} (A_T v\,|\,P_s w_j) K_T g_j \right\|_0 \\ &= \left\| \sum_{j=1}^{N} (v\,|\,A_T^* P_s w_j) K_T g_j \right\|_0 \\ &\leqq \|v\|_0 \sum_{j=1}^{N} \|A_T^* P_s w_j\|_0 \|K_T g_j\|_0. \end{aligned}$$

This shows that $K_T T'' A_T$ acts as an L^2-bounded operator on $D(A_T)$, when $P_s w_j \in D(A_T^*)$, $j = 1, 2, \cdots, N$.

Therefore, (5.94) (and with it (5.92) in a related sense) is simply a *bounded* perturbation of

$$(5.95)\qquad \begin{aligned} &\partial_t^2 v + Av = 0 && \text{in } \Omega \quad \text{for } t \in \mathbb{R}, \\ &Tv = 0 && \text{on } \Gamma \quad \text{for } t \in \mathbb{R}, \\ &v = v_0 && \text{in } \Omega \quad \text{at } t = 0, \\ &\partial_t v = v_1 && \text{in } \Omega \quad \text{at } t = 0, \end{aligned}$$

treated in Theorem 5.6. Since the spectrum of A_T is assumed to be contained in $\mathbb{R}_+$, we find from standard bounded perturbation theory (see, e.g., Sova [12] and Fattorini [3]) that the operators

$$\cos((A_T - K_T T'' A_T)t), \tag{5.96}$$

$$\sin((A_T - K_T T'' A_T)t) \tag{5.97}$$

are well-defined, bounded, bounded operators in $L^2(\Omega)$ for $P_s w_j \in D(A_T^*)$, $j = 1, 2, \cdots, N$, and $t \in \mathbb{R}$. From Theorem 1.6.11 of Grubb [5] we find that

$$D(A_T^*) = \{u \in H^{2m}(\Omega) \,|\, \dot{I}^x \mathscr{A}^{01*} \gamma u = 0\} \tag{5.98}$$

where $\dot{I}^x$ is the "reflection" of the index set, replacing $\{k\}_{0 \leqq k < m}$ by $\{2m-k-1\}_{0 \leqq k < m}$, and $\mathscr{A}^{01}$ is the $m \times m$ matrix differential operator, appearing in Green's formula ((A4) in the Appendix) for A, it is invertible since A is elliptic. We see from Example 1.6.12. of Grubb [5] that the action of the realization A_T^* is of the form $A + G$, for a certain singular Green operator G of finite rank, whereas the domain $D(A_T^*)$ is simply characterized as

$$D(A_T^*) = D(A_\gamma) = H^{2m}(\Omega) \cap H_0^m(\Omega). \tag{5.99}$$

We have thus obtained Theorem 5.11.

THEOREM 5.11. *Let the set* $\{w_j, g_j\}_{1 \leqq j \leqq N}$ *be chosen according to Theorem* 5.10, *and assume furthermore that* $P_s w_j$, $j = 1, 2, \cdots, N$, *are chosen in* $D(A_\gamma) = H^{2m}(\Omega) \cap H_0^m(\Omega)$. *Then the operators*

$$C(t) = \cos((A_T - K_T T'' A_T)t) \tag{5.100}$$

and

$$S(t) = (A_T - K_T T'' A_T)^{-1/2} \sin((A_T - K_T T'' A_T)^{1/2} t) \tag{5.101}$$

on $L^2(\Omega)$, *are well defined for* $t \in \mathbb{R}$, *giving the solution to the system*

$$\begin{aligned} &\partial_t^2 v + (1 - K_T T'')Av = 0 && \textit{in } \Omega \quad \textit{for } t \in \mathbb{R}, \\ &\gamma v = \sum_{j=1}^{N} (v \,|\, P_u w_j) g_j && \textit{on } \Gamma \quad \textit{for } t \in \mathbb{R}, \\ &v = v_0 && \textit{in } \Omega \quad \textit{at } t = 0, \\ &\partial_t v = v_1 && \textit{in } \Omega \quad \textit{at } t = 0, \end{aligned} \tag{5.102}$$

as

$$v(t, x) = C(t)v_0(x) + S(t)v_1(x), \qquad x \in \Omega, \quad t \in \mathbb{R}, \quad v_0, v_1 \in L^2(\Omega). \tag{5.103}$$

The solution to the original system

$$\begin{aligned} &\partial_t^2 u + Au = 0 && \textit{in } \Omega \quad \textit{for } t \in \mathbb{R}, \\ &\gamma u = \sum_{j=1}^{N} (u \,|\, w_j) g_j && \textit{on } \Gamma \quad \textit{for } t \in \mathbb{R}, \\ &u = u_0 && \textit{in } \Omega \quad \textit{at } t = 0, \\ &\partial_t u = u_1 && \textit{in } \Omega \quad \textit{at } t = 0, \end{aligned} \tag{5.104}$$

is then by (5.74)

$$\begin{aligned} u(t, x) = {} & (1 - K_T T'')^{-1} C(t)(1 - K_T T'')u_0(x) \\ & + (1 - K_T T'')^{-1} S(t)(1 - K_T T'')u_1(x), \ x \in \Omega, \quad t \in \mathbb{R}, \quad u_0, u_1 \in L^2(\Omega). \end{aligned} \tag{5.105}$$

6. Stabilization by perturbations of the third kind. To get a complete picture, we shall now also discuss briefly how we can stabilize the systems (1.1) and (1.2) by changing *both* the boundary condition and the system operator. The stabilized systems will then take the form

$$\begin{aligned} &\partial_t u + Au + Gu = 0 && \text{in } \Omega && \text{for } t \geqq 0, \\ &\gamma u = T'u && \text{on } \Gamma && \text{for } t \geqq 0, \\ &u = u_0 && \text{in } \Omega && \text{at } t = 0, \end{aligned} \tag{6.1}$$

respectively,

$$\begin{aligned} &\partial_t^2 u + Au + Gu = 0 && \text{in } \Omega && \text{for } t \in \mathbb{R}, \\ &\gamma u = T'u && \text{on } \Gamma && \text{for } t \in \mathbb{R}, \\ &u = u_0 && \text{in } \Omega && \text{at } t = 0, \\ &\partial_t u = u_1 && \text{in } \Omega && \text{at } t = 0, \end{aligned} \tag{6.2}$$

and we say that the systems (6.1) and (6.2) are associated with the realization $B = (A+G)_T$, where

$$Bu = (A+G)u, \tag{6.3}$$

defined on

$$D(B) = \{u \in H^{2m}(\Omega) \mid Tu = 0\}, \tag{6.4}$$

with

$$T = \gamma - T'. \tag{6.5}$$

We will now determine operators G and T', such that (6.1) and (6.2) are stable systems of feedback type. Here, we can, in fact, make B a self-adjoint generator of an analytic *contraction* semigroup. Here we will point out that the following construction is not very "economical," as we typically use a large number of feedback terms, but the construction clarifies to a great extent the interaction between the system operator equation and the boundary equation.

As in the preceding paragraph, let $\{\varphi_j\}_{j \geqq 1}$ be an orthonormalized set of eigenfunctions for A_γ, enumerated according to the ordering of the eigenvalues (5.5), but now define the boundary feedback operator T' as

$$T'u = \sum_{j=1}^{K-1} (u \mid \varphi_j) h_j, \tag{6.6}$$

where $K-1$ is the number of negative eigenvalues of A_γ, repeated according to multiplicity, and

$$h_j \in C^\infty(\Gamma)^m, \qquad j = 1, 2, \cdots, K-1, \tag{6.7}$$

are chosen linearly independent, with no other conditions on them. Moreover, define the operator G as

$$G = K'\nu + G', \tag{6.8}$$

where ν is the Neumann trace operator (1.6) and

$$K' = -T'^* \mathscr{A}^{01}, \qquad G' = -T'^*(\mathscr{S} - c)\gamma. \tag{6.9}$$

Here c is a positive constant to be determined later, $\mathscr{A}^{01}$ is the upper right corner in the coefficient matrix (A3) in Green's formula (see the Appendix), whereas $\mathscr{S}$ is the coefficient matrix appearing in the boundary term in the "halfways" Green formula for a convenient sesquilinear form $a(u, v)$ associated with A (see (A6)).

Let $a(u, v)$ be a symmetric sesquilinear form on $H^m(\Omega)$ associated with A, of the form (A5), and recall that, by the Gårding inequality, $a(u, v)$ is $H_0^m(\Omega)$-coercive. It is well known how A_γ is the *variational operator* associated with the triple $(a, H_0^m(\Omega), L^2(\Omega))$ (see, e.g., Grubb [5, § 1.7]). Let us define the sesquilinear form $a_1(u, v)$ on $H^m(\Omega)$ by

$$a_1(u, v) = a(u, v) + c(T'u \mid \gamma v)_\Gamma. \tag{6.10}$$

Let U be the closed subspace of $H^m(\Omega)$:

$$U = \{u \in H^m(\Omega) \mid \gamma u = T'u\}, \tag{6.11}$$

and observe that U is dense in $L^2(\Omega)$, since $T = \gamma - T'$ defines a normal boundary condition $Tu = 0$ (so that already $D(A_T)$ is dense in $L^2(\Omega)$; cf. Lemma 3.1). Let B_1 be the operator associated with the triple $(a_1, U, L^2(\Omega))$, defined as follows:

$$\begin{aligned} &D(B_1) = \{u \in U \mid \exists f \in L^2(\Omega) \text{ so that } a_1(u, v) = (f \mid v) \text{ for all } v \in U\}, \\ &B_1 u = f. \end{aligned} \tag{6.12}$$

We will show that $B = B_1$, where B is the realization defined in (6.3)-(6.5).

LEMMA 6.1.

$$B \subseteq B_1. \tag{6.13}$$

Proof. For $u \in D(B)$, $v \in U$ we have in view of (6.4), (6.9), (6.11), and formula (A6) of the Appendix:

$$\begin{aligned} a_1(u, v) &= a(u, v) + c(T'u \mid \gamma v)_\Gamma \\ &= (Au \mid v) - (\mathscr{A}^{01}\nu u + \mathscr{S}\gamma u \mid \gamma v)_\Gamma + c(T'u \mid \gamma v)_\Gamma \\ &= (Au \mid v) - (\mathscr{A}^{01}\nu u + \mathscr{S}\gamma u \mid T'v)_\Gamma + c(\gamma u \mid T'v)_\Gamma \\ &= (Au - T'^*(\mathscr{S} - c)\gamma u - T'^*\mathscr{A}^{01}\nu u \mid v) \\ &= ((A + G)u \mid v). \end{aligned}$$

This shows that $u \in D(B_1)$ with $B_1 u = (A + G)u$, so it follows that $D(B) \subset D(B_1)$ with $B_1 u = Bu$ there. □

For $l \in \mathbb{N}$ we define the subspace W_l of $H_0^m(\Omega)$ by

$$W_l = \overline{\operatorname{span}}^m \{\varphi_j\}_{j \leqq l} \qquad (H^m(\Omega) - \text{closure}). \tag{6.14}$$

Note that for $u \in W_K$, γu and $T'u$ are zero, so $W_K \subseteq U$.

LEMMA 6.2. *There exists a linearly independent set of functions $\{v_j\}_{1 \leqq j < K}$ in $U \setminus W_K$, such that*

$$U = \operatorname{span}\{v_j\}_{1 \leqq j < K} \dotplus W_K. \tag{6.15}$$

Proof. According to Lemma 1.6.8 of Grubb [5] (see our § 3) we can write U of the form

$$U = \Lambda^{-1} H_0^m(\Omega) \tag{6.16}$$

where Λ and Λ^{-1} are bounded operators in $H^s(\Omega)$, for all $s \geqq 0$.

Choose a linearly independent set $\{z_j\}_{1\leqq j<K}$ in $H_0^m(\Omega)$ such that the matrix

$$C=((z_i\,|\,(\Lambda^{-1})^*\varphi_j)_{i,j})_{1\leqq i,j<K} \tag{6.17}$$

is regular, and define

$$v_j=\Lambda^{-1}z_j, \qquad j=1,2,\cdots,K-1; \tag{6.18}$$

this is clearly a linearly independent set in U.

Moreover, for $1\leqq j<K$

$$\begin{aligned}\gamma v_j=\gamma\Lambda^{-1}z_j&=\sum_{k=1}^{K-1}(\Lambda^{-1}z_j\,|\,\varphi_k)h_k\\&=\sum_{k=1}^{K-1}(z_j\,|\,(\Lambda^{-1})^*\varphi_k)h_k\neq 0\end{aligned} \tag{6.19}$$

since C is regular, so none of the v_j lie in $H_0^m(\Omega)$.

Since

$$\gamma\begin{pmatrix}v_1\\v_2\\\vdots\\v_{K-1}\end{pmatrix}=C\begin{pmatrix}h_1\\h_2\\\vdots\\h_{K-1}\end{pmatrix} \tag{6.20}$$

the set $\{\gamma v_j\}_{1\leqq j<K}$ is linearly independent, so $\gamma(\sum_{j=1}^{K-1}\alpha_jv_j)=0$ implies that $\alpha_j=0$, $j=1,2,\cdots,K-1$. Then also

$$\operatorname{span}\{v_j\}_{1\leqq j<K}\cap W_K=\{0\}.$$

Since $W_K\subseteq U$ and $\operatorname{span}\{v_j\}_{1\leqq j<K}\subseteq U$, the inclusion

$$\operatorname{span}\{v_j\}_{1\leqq j<K}\dotplus W_K\subseteq U \tag{6.21}$$

is evident.

To show the inclusion the other way, let $u\in U$ and invert (6.20) to get

$$\gamma u=\sum_{j=1}^{K-1}(u\,|\,\varphi_j)h_j=\sum_{j=1}^{K-1}\left((u\,|\,\varphi_j)\sum_{k=1}^{K-1}\alpha_{jk}\gamma v_k\right), \qquad \alpha_{jk}\in\mathbb{C}. \tag{6.22}$$

The last term equals $\sum_{k=1}^{K-1}\beta_k\gamma v_k$, where

$$\beta_k=\sum_{j=1}^{K-1}(u\,|\,\varphi_j)\alpha_{jk}, \qquad k=1,2,\cdots,K-1. \tag{6.23}$$

If we define

$$w=u-\sum_{j=1}^{K-1}\beta_jv_j \tag{6.24}$$

then by (6.22),

$$\gamma w=\gamma u-\sum_{j=1}^{K-1}\beta_j\gamma v_j=0, \tag{6.25}$$

hence (since u and the v_j lie in U)

$$0=\gamma w=\sum_{j=1}^{K-1}(w\,|\,\varphi_j)h_j. \tag{6.26}$$

But the set $\{h_j\}_{1\leqq j<K}$ is linearly independent, so (6.26) implies that

$$(w\,|\,\varphi_j)=0, \qquad j=1,2,\cdots,K-1. \tag{6.27}$$

Then, from (6.24)

$$u=\sum_{j=1}^{K-1}\beta_j v_j+w\in \operatorname{span}\{v_j\}_{1\leqq j<K}\dotplus W_K. \qquad \square \tag{6.28}$$

Since $\bar{W}_K^0$ (i.e., the $L^2(\Omega)$-closure) has a L^2-complement $(\bar{W}_K^0)^\perp$ of dimension $K-1$, we can obtain (by a small perturbation, if necessary) that the v_j are chosen outside $\bar{W}_K^0$. More precisely, the following construction can be used:

Let $v_j=\varphi_j+a_j$, where the a_j are chosen such that

(1) $v_j\notin \bar{W}^0$. This is achieved if $\|a_j\|_0\leqq\frac{1}{2}$, since $\|\varphi_j\|_0=1$ and $\varphi_j\perp W_K$.

(2) $C=((z_i\,|\,(\Lambda^{-1})^*\varphi_j)_{i,j})_{1\leqq i,j<K}$ is regular, when $z_i=\Lambda v_i$. Since $(z_i\,|\,(\Lambda^{-1})^*\varphi_j)=(\varphi_i+a_i\,|\,\varphi_j)=\delta_{ij}+(a_i\,|\,\varphi_j)$, this is achieved by choosing $\max_j\|a_j\|_0$ so small that the norm of the matrix $((a_i\,|\,\varphi_j)_{i,j})_{1\leqq i,j<K}$ is $\leqq\frac{1}{2}$.

(3) $z_j\in H_0^m(\Omega)$. Since $C_0^\infty(\Omega)$ is dense in $L^2(\Omega)$ we can, for a given $\varepsilon>0$ choose $z_j\in C_0^\infty(\Omega)$ such that $\|z_j-\Lambda\varphi_j\|_0\leqq\varepsilon$. Then $a_j=\Lambda^{-1}(z_j-\Lambda\varphi_j)$, with norm $\|a_j\|_0\leqq \varepsilon\|\Lambda^{-1}\|_{L^2,L^2}$. If ε is then adjusted so that (1) and (2) hold, we have a solution $\{v_j\}_{1\leqq j<K}$ with the required property. Then we have a decomposition of $u\in U$ as

$$u=v+w \tag{6.29}$$

where $v\in\operatorname{span}\{v_j\}_{1\leqq j<K}$, $v_j\notin\bar{W}_K^0 j=1,2,\cdots,K-1$, and $w\in W_K$, and the projections

$$u\to v, \qquad u\to w \tag{6.30}$$

are then continuous in $L^2(\Omega)$ so we have an inequality

$$\|v\|_0+\|w\|_0\leqq C\|u\|_0 \tag{6.31}$$

with a constant $C>0$.

LEMMA 6.3. *$a(u,v)$ is m-coercive on U, i.e., (since a is symmetric on $H^m(\Omega)$),*

$$\operatorname{Re} a(u,u)=a(u,u)\geqq C_0\|u\|_m^2-k\|u\|_0^2 \tag{6.32}$$

for $u\in U$, with constants $C_0>0$ and $k\in\mathbb{R}$.

Proof. Recall that the L^2-norm and the H^m-norm are equivalent on a finite-dimensional space, so that we have positive constants C' and C'', such that

$$C'\|v\|_0\leqq\|v\|_m\leqq C''\|v\|_0. \tag{6.33}$$

Let $u\in U$ and write $u=v+w$ as in (6.29). Then

$$\operatorname{Re} a(u,u)=a(u,u)=a(w,w)+2\operatorname{Re} a(v,w)+a(v,v).$$

For any $\varepsilon>0$ we can find constants C_1, $C_2>0$, such that

$$\begin{aligned}2|a(v,w)|&\leqq 2C_1\|w\|_m\|v\|_m\\ &\leqq\varepsilon^2\|w\|_m^2+\frac{C_1^2}{\varepsilon^2}\|v\|_m^2\\ &\leqq\varepsilon^2\|w\|_m^2+\frac{C_2}{\varepsilon^2}\|u\|_0^2.\end{aligned} \tag{6.34}$$

Here we have used (6.31) for v, together with (6.33).

Since $a(u, v)$ is $H_0^m(\Omega)$-coercive, we find, using (6.31) for w,

$$
\begin{aligned}
a(w, w) &\geqq C_3\|w\|_m^2 - C_4\|w\|_0^2 \\
&\geqq C_3\|w\|_m^2 - C_5\|u\|_0^2,
\end{aligned}
\tag{6.35}
$$

with constants $C_3, C_4, C_5 > 0$.

Moreover, since

$$
a(v, v) \leqq C_6\|v\|_m^2 \leqq C_7\|v\|_0^2 \leqq C_8\|u\|_0^2,
\tag{6.36}
$$

with $C_6, C_7, C_8 > 0$, we have that

$$
a(u, u) \geqq (C_3 - \varepsilon^2)\|w\|_m^2 - \left(\frac{C_2}{\varepsilon^2} + C_5 + C_8\right)\|u\|_0^2.
\tag{6.37}
$$

Then, for all $\varepsilon' > 0$:

$$
\begin{aligned}
\|w\|_m^2 = \|u - v\|_m^2 &\geqq \|u\|_m^2 - 2\|v\|_m\|u\|_m + \|v\|_m^2 \\
&\geqq \|u\|_m^2 + C'^2\|v\|_0^2 - \left(\frac{1}{\varepsilon'^2}\|v\|_m^2 + \varepsilon'^2\|u\|_m^2\right) \\
&\geqq (1 - \varepsilon'^2)\|u\|_m^2 + \left(C'^2 - \frac{1}{\varepsilon'^2}C''^2\right)\|v\|_0^2 \\
&\geqq (1 - \varepsilon'^2)\|u\|_m^2 - \left|C'^2 - \frac{1}{\varepsilon'^2}C''^2\right| C\|u\|_0^2.
\end{aligned}
\tag{6.38}
$$

Choosing $0 < \varepsilon^2 < C_3$ and $0 < \varepsilon' < 1$ and inserting the result of (6.38) in (6.37), we finally obtain (6.32). □

LEMMA 6.4. $B_1 = B_1^*$.

Proof. Since for $u \in U$

$$
a_1(u, u) = a(u, u) + c(T'u \mid \gamma u)_\Gamma = a(u, u) + c(T'u \mid T'u)_\Gamma \geqq a(u, u),
\tag{6.39}
$$

a_1 is m-coercive on U. Then B_1 is the *variational* operator associated with the triple $(a_1, U, L^2(\Omega))$, and since $a_1(u, v)$ is symmetric on U, $B_1 = B_1^*$. (see, e.g., Grubb [5], § 1.7.) □

Now we note that the boundary value problem

$$
\begin{aligned}
&Au + Gu = f, \qquad f \in L^2(\Omega), \\
&Tu = 0
\end{aligned}
\tag{6.40}
$$

is elliptic, since the Dirichlet problem for A

$$
\begin{aligned}
&Au = f, \qquad f \in L^2(\Omega), \\
&\gamma u = 0
\end{aligned}
\tag{6.41}
$$

is elliptic and G has the form (6.8)-(6.9) where T'^* and T' obviously are integral operators with C^∞-kernels, hence of order $-\infty$. (Problem (6.40) and (6.41) have the same principal symbols.) We can then use Theorem 1.6.11 and Example 1.6.12 of Grubb [5] to show that $B = B^*$. It is shown there that when

$$
B = (A + G)_T
\tag{6.42}
$$

is an elliptic realization of a formally self-adjoint operator A, with

$$
G = K'\nu + G', \qquad T = \gamma - T'
\tag{6.43}
$$

then $B=B^*$ if

$$T'=-(\mathscr{A}^{01^*})^{-1}K^{1^*},\qquad G'=G'^*+T'^*\mathscr{A}^{00^*}\gamma. \tag{6.44}$$

Using that $K'=-T'^*\mathscr{A}^{01}$ (see (6.9)) and that $\mathscr{A}^{00}=\mathscr{S}-\mathscr{S}^*$ (see Lemma A1 of the appendix), we compute

$$-(\mathscr{A}^{01*})^{-1}K'^*=(\mathscr{A}^{01*})^{-1}\mathscr{A}^{01*}T'=T' \tag{6.45}$$

and

$$\begin{aligned} G'^*+T'^*\mathscr{A}^{00}\gamma &= -T'^*(\mathscr{S}^*-c)\gamma+T'^*(\mathscr{S}^*-\mathscr{S})\gamma \\ &= -T'^*(\mathscr{S}-c)\gamma \\ &= G', \end{aligned} \tag{6.46}$$

hence $B=B^*$, according to (6.44).

We now have Theorem 6.5.

THEOREM 6.5.

$$B=B_1. \tag{6.47}$$

Proof. $B\subseteq B_1$ implies that $B_1^*\subseteq B^*$. But $B=B^*$ and $B_1=B_1^*$. □

We have hereby shown that the realization

$$B=(A+G)_T \tag{6.48}$$

is the *variational* operator associated with the triple $(a_1, U, L^2(\Omega))$.

THEOREM 6.6. *Assume that the operators G and T are chosen such that the operator $B=(A+G)_T$ is the variational operator associated with the triple $(a_1, U, L^2(\Omega))$, and let $\zeta<1$ be given. There exists a constant $c>0$ such that the sesquilinear form*

$$a_1(u,v)=a(u,v)+c(T'u\,|\,\gamma v)_\Gamma \tag{6.49}$$

satisfies, for all $u\in U$

$$a_1(u,u)\geqq \zeta\lambda_K\|u\|_0^2. \tag{6.50}$$

Proof. We write $u\in U$ as $u=v+w$ as in (6.29). Then

$$\begin{aligned} \frac{a_1(u,u)}{\|u\|_0^2} &= \frac{a(v+w,v+w)+c(T'(v+w)\,|\,\gamma(v+w))_\Gamma}{(v+w\,|\,v+w)} \\ &= \frac{a(v+w,v+w)+c(T'(v+w)\,|\,T'(v+w))_\Gamma}{(v+w\,|\,v+w)} \\ &= \frac{a(w,w)+2\,\mathrm{Re}\,a(v,w)+a(v,v)+c\|T'v\|_0^2}{\|w\|_0^2+2\,\mathrm{Re}\,(v\,|\,w)+\|v\|_0^2}. \end{aligned}$$

Now since

$$\begin{aligned} &a(w,w)\geqq C_1\|w\|_m^2-C_2\|w\|_0^2, \qquad C_1, C_2>0, \\ &a(w,w)\geqq \lambda_K\|w\|_0^2 \end{aligned} \tag{6.51}$$

we have, for all $\varepsilon > 0$

$$(6.52)\quad \begin{aligned} 2|a(v, w)| &\leqq \varepsilon^2\|w\|_m^2 + \frac{C_3}{\varepsilon^2}\|v\|_m^2 \\ &\leqq \varepsilon^2\|w\|_m^2 + \frac{C_4}{\varepsilon^2}\|v\|_0^2 \\ &\leqq \varepsilon^2\left(\frac{1}{C_1}a(w, w) + \frac{C_2}{C_1}\|w\|_0^2\right) + \frac{C_4}{\varepsilon^2}\|v\|_0^2 \\ &\leqq \varepsilon^2\left(\frac{1}{C_1} + \frac{1}{\lambda_K}\cdot\frac{C_2}{C_1}\right)a(w, w) + \frac{C_4}{\varepsilon^2}\|v\|_0^2 \end{aligned}$$

with constants $C_3, C_4 > 0$.

Moreover, for all $\varepsilon' > 0$

$$(6.53)\quad 2|(v\,|\,w)| \leqq 2\|v\|_0\|w\|_0 \leqq \varepsilon'^2\|w\|_0^2 + \frac{1}{\varepsilon'^2}\|v\|_0^2,$$

so we find that

$$(6.54)\quad \frac{a_1(u, u)}{\|u\|_0^2} \geqq \frac{\left(1 - \varepsilon^2\left(\frac{1}{C_1} + \frac{1}{\lambda_K}\cdot\frac{C_2}{C_1}\right)\right)a(w, w) - \left(\frac{C_4}{\varepsilon^2} + C_6\right)\|v\|_0^2 + c\|T'v\|_0^2}{(1 + \varepsilon'^2)\|w\|_0^2 + \left(1 + \frac{1}{\varepsilon'^2}\right)\|v\|_0^2},$$

where we also used that $|a(v, v)| \leqq C_5\|v\|_m^2 \leqq C_6\|v\|_0^2$, $C_5, C_6 > 0$ (cf. (6.33)).

Then, in particular,

$$(6.55)\quad \frac{a_1(u, u)}{\|u\|_0^2} \geqq \frac{\left(1 - \varepsilon^2\left(\frac{1}{C_1} + \frac{1}{\lambda_K}\cdot\frac{C_2}{C_1}\right)\right)\lambda_K\|w\|_0^2 - \left(\frac{C_4}{\varepsilon^2} + C_6\right)\|v\|_0^2 + c\|T'v\|_0^2}{(1 + \varepsilon'^2)\|w\|_0^2 + \left(1 + \frac{1}{\varepsilon'^2}\right)\|v\|_0^2}.$$

T' is injective from span $\{v_j\}_{1\leqq j<K}$ to span $\{h_j\}_{1\leqq j<K}$, so we have that

$$(6.56)\quad \|T'v\|_0^2 \geqq C_7\|v\|_0^2,$$

with a positive constant C_7.

Hence

$$(6.57)\quad \frac{a_1(u, u)}{\|u\|_0^2} \geqq \frac{\alpha\lambda_K\|w\|_0^2 + \beta\|v\|_0^2}{\mu\|w\|_0^2 + \theta\|v\|_0^2},$$

where

$$\alpha = 1 - \varepsilon^2\left(\frac{1}{C_1} + \frac{1}{\lambda_K}\cdot\frac{C_2}{C_1}\right), \qquad \beta = cC_7 - \frac{C_4}{\varepsilon^2} - C_6,$$

$$\mu = 1 + \varepsilon'^2, \qquad \theta = 1 + \frac{1}{\varepsilon'^2}.$$

Considering the function

$$(6.58)\quad f(s) = \frac{\alpha\lambda_K s + \beta}{\mu s + \theta}, \qquad s \geqq 0,$$

which is nonincreasing for $s\to\infty$ when $\lambda_K\alpha\theta-\mu\beta\leqq 0$, we see that

$$\frac{a_1(u,u)}{\|u\|_0^2}\geqq\frac{\alpha}{\mu}\lambda_K, \tag{6.59}$$

if

$$\beta\geqq\frac{\alpha\theta}{\mu}\lambda_K. \tag{6.60}$$

Then if we choose c in the definition of a_1 such that

$$c\geqq\frac{\left(1-\varepsilon^2\left(\frac{1}{C_1}+\frac{1}{\lambda_K}\cdot\frac{C_2}{C_1}\right)\right)\lambda_K\left(1+\frac{1}{\varepsilon'^2}\right)}{C_7(1+\varepsilon'^2)}+\frac{C_4}{C_7\varepsilon^2}+\frac{C_6}{C_7}, \tag{6.61}$$

then

$$a_1(u,u)\geqq\frac{1-\varepsilon^2\left(\frac{1}{C_1}+\frac{1}{\lambda_K}\cdot\frac{C_2}{C_1}\right)}{1+\varepsilon'^2}\lambda_K\|u\|_0^2. \tag{6.62}$$

With ε and ε' chosen such that

$$\frac{1-\varepsilon^2\left(\frac{1}{C_1}+\frac{1}{\lambda_K}\cdot\frac{C_2}{C_1}\right)}{1+\varepsilon'^2}=\zeta$$

and c chosen such that (6.61) holds, we see that

$$a_1(u,u)\geqq\zeta\lambda_K\|u\|_0^2 \tag{6.63}$$

as claimed. □

We now have Theorem 6.7.

THEOREM 6.7. *With the hypotheses of Theorem* 6.6 *we have that, given any* $\zeta<1$, *there exists a constant* $c>0$ *such that the self-adjoint operator realization*

$$B=(A+G)_T\quad (\textit{see } (6.3)\text{-}(6.5)) \tag{6.64}$$

has its spectrum in the halfline $[\zeta\lambda_K,\infty[$, *and is the infinitesimal generator of an analytic semigroup* e^{-Bt}, $t\geqq 0$, *on* $L^2(\Omega)$, *giving the solutions to the parabolic system*

$$\begin{aligned} &\partial_t u+Au+Gu=0 && \textit{in } \Omega && \textit{for } t>0,\\ &\gamma u=T'u && \textit{on } \Gamma && \textit{for } t>0,\\ &u=u_0 && \textit{in } \Omega && \textit{at } t=0, \end{aligned} \tag{6.65}$$

as

$$u(t,x)=e^{-Bt}u_0(x),\qquad x\in\Omega,\quad t\geqq 0,\quad u_0\in L^2(\Omega). \tag{6.66}$$

The solution satisfies

$$\|u(t,\cdot)\|_0\leqq e^{-\zeta\lambda_K t}\|u_0\|_0,\qquad t\geqq 0. \tag{6.67}$$

Moreover, the operators

$$\begin{aligned} C(t)&=\cos(B^{1/2}t),\\ S(t)&=B^{-1/2}\sin(B^{1/2}t) \end{aligned} \tag{6.68}$$

are well defined for $t \in \mathbb{R}$, giving the solution to the hyperbolic problem

$$(6.69)\qquad \begin{array}{lll} \partial_t^2 u + Au + Gu = 0 & \text{in } \Omega & \text{for } t \in \mathbb{R}, \\ \gamma u = T'u & \text{on } \Gamma & \text{for } t \in \mathbb{R}, \\ u = u_0 & \text{in } \Omega & \text{at } t = 0, \\ \partial_t u = u_1 & \text{in } \Omega & \text{at } t = 0, \end{array}$$

as

$$(6.70)\qquad u(t,x) = C(t)u_0(x) + S(t)u_1(x), \qquad x \in \Omega, \quad t \in \mathbb{R},$$

when $u_0, u_1 \in L^2(\Omega)$.

Remark 6.8. In the light of Proposition 1.7.11 of Grubb [5], it is of interest to note that the realization constructed above is weakly semibounded and m-bounded (cf. Remark 5.7) because we have compensated for the nonlocalness in the boundary condition by nonlocal terms in the system operator equation, satisfying

$$(6.71)\qquad K'^* = -\mathscr{A}^{01*}T'.$$

Example 6.9. Let us calculate the operator G in the case where $A = -\Delta$. Since

$$(6.72)\qquad T'u = \sum_{j=1}^{K-1} (u \mid \varphi_j)h_j,$$

we have that

$$(6.73)\qquad T'^*\psi = \sum_{j=1}^{K-1} (\psi \mid h_j)_\Gamma \varphi_j,$$

hence

$$(6.74)\qquad \begin{aligned} Gu &= -T'^*\mathscr{A}^{01}\nu u - T'^*(\mathscr{S} - c)\gamma u \\ &= -\sum_{j=1}^{K-1} (\mathscr{A}^{01}\nu u - (\mathscr{S} - c)\gamma u \mid h_j)_\Gamma \varphi_j. \end{aligned}$$

Now the terms in the Green formula (see (A4) of the appendix) are particularly simple since

$$(6.75)\qquad \mathscr{A} = i\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},$$

$$(6.76)\qquad (-\Delta u \mid v) - (u \mid -\Delta v) = i(\nu u \mid \gamma v)_\Gamma + i(\gamma u \mid \nu v)_\Gamma.$$

Then (6.74) reduces to

$$(6.77)\qquad Gu = -\sum_{j=1}^{K-1} (i\nu u + c\gamma u \mid h_j)_\Gamma \varphi_j.$$

Appendix. Green's formulas. Using the notation given in (1.5)-(1.8) we have, for the formally self-adjoint operator A (1.3)

$$(\mathrm{A1})\qquad (Au \mid v) - (u \mid Av) = (\mathscr{A}\rho u \mid \rho v)_\Gamma,$$

where $\mathscr{A}$ is a skew-triangular $2m\times 2m$ matrix of differential operators over the boundary Γ, of the form

$$\text{(A2)}\qquad \mathscr{A}=\begin{pmatrix} S_1^0 & \cdots & S_{2m-1}^0 & S_{2m}^0\\ S_2^0 & \cdots & S_{2m}^0 & 0\\ \vdots & & & \\ S_{2m}^0 & \cdots & 0 & 0\end{pmatrix}+\begin{pmatrix}\text{lower order} & & 0\\ & \cdot^{\cdot^{\cdot}} & \vdots\\ 0 & \cdots & 0\end{pmatrix},$$

the S_k^0 being differential operators on Γ of order $2m-k$ (see, e.g., Grubb [5], § 1.3).

We usually write $\mathscr{A}$ in $m\times m$ blocks as

$$\text{(A3)}\qquad \mathscr{A}=\begin{pmatrix}\mathscr{A}^{00} & \mathscr{A}^{01}\\ \mathscr{A}^{10} & 0\end{pmatrix}$$

and we have the following version of the Green formula:

$$\text{(A4)}\qquad (Au\,|\,v)-(u\,|\,Av)=(\mathscr{A}^{01}\nu u+\mathscr{A}^{00}\gamma u\,|\,\gamma v)_\Gamma+(\mathscr{A}^{10}\gamma u\,|\,\nu v)_\Gamma$$

for $u, v\in H^{2m}(\Omega)$. The coefficient matrices $\mathscr{A}^{ij}$ here are uniquely determined from A.

Now consider a *symmetric* sesquilinear form with $a_{\alpha\beta}\in C^\infty(\bar\Omega)$, $\bar a_{\alpha\beta}=a_{\beta\alpha}$ for all α, β,

$$\text{(A5)}\qquad a(u,v)=\sum_{|\alpha|,|\beta|\leqq m}(a_{\alpha\beta}D^\beta u\,|\,D^\alpha v)$$

associated with A. For such a form we have a "halfways" Green formula, for $u\in H^{2m}(\Omega)$ and $v\in H^m(\Omega)$:

$$\text{(A6)}\qquad (Au\,|\,v)-a(u,v)=(\mathscr{A}^{01}\nu u+\mathscr{S}\gamma u\,|\,\gamma v)_\Gamma$$

where the operator $\mathscr{S}$ is of the same type as $\mathscr{A}^{00}$ in (A3). Since $\mathscr{A}^{01^*}=-\mathscr{A}^{10}$ when A is formally self-adjoint, we have Lemma A1.

LEMMA A1.

$$\text{(A7)}\qquad \mathscr{A}^{00}=\mathscr{S}-\mathscr{S}^*.$$

Proof. For $u, v\in H^{2m}(\Omega)$ we have

$$\begin{aligned}(Au\,|\,v)&=a(u,v)+(\mathscr{A}^{01}\nu u+\mathscr{S}\nu u\,|\,\gamma v)_\Gamma,\\ (u\,|\,Av)=\overline{(Av\,|\,u)}&=\overline{a(u,v)+(\mathscr{A}^{01}\nu v+\mathscr{S}\gamma v\,|\,\gamma u)_\Gamma}\\ &=a(u,v)+(\mathscr{A}^{01*}\gamma u\,|\,\nu v)_\Gamma+(\mathscr{S}^*\gamma u\,|\,\gamma v)_\Gamma\end{aligned}$$

so that

$$\text{(A8)}\qquad (Au\,|\,v)-(u\,|\,Av)=(\mathscr{A}^{01}\nu u+(\mathscr{S}-\mathscr{S}^*)\gamma u\,|\,\gamma v)_\Gamma+(\mathscr{A}^{10}\gamma u\,|\,\nu v)_\Gamma.$$

Comparing (A8) with (A4) gives us (A7). □

Now let K_γ be the Poisson solution operator to the Dirichlet problem for A, i.e., K_γ maps ψ into u, where

$$\text{(A9)}\qquad Au=0\quad\text{in }\Omega,\qquad \gamma u=\psi\quad\text{on }\Gamma.$$

We can then specify the action of K_γ the following way.

PROPOSITION A2. *The Poisson solution operator K_γ to the Dirichlet problem for A satisfies*

$$\text{(A10)}\qquad (K_\gamma\psi\,|\,\varphi_j)=\frac{-1}{\lambda_j}(\psi\,|\,\mathscr{A}^{10^*}\nu\varphi_j)_\Gamma,\qquad j\geqq 1.$$

Proof. Just insert $u = K_\gamma \psi$ and $v = \varphi_j$ (the eigenfunctions of A_γ) in (A4), and use that $\gamma v = 0$. Moreover, $\lambda = \bar{\lambda}$ since A is self-adjoint. □

Acknowledgments. The author is deeply grateful to Gerd Grubb, University of Copenhagen, for her patience in teaching us the pseudodifferential theory, and to Martin Bendsøe, Technical University of Denmark, for many fruitful discussions and much advice.

REFERENCES

[1] A. V. BALAKRISHNAN, *Boundary control of parabolic equations*: L-Q-R-*Theory*. in Proc. Conference on Theory of Nonlinear Equations, September 1977, Akademie Verlag, Berlin, 1978.

[2] L. BOUTET DE MONVEL, *Boundary problems for pseudo-differential operators*, Acta Math., 126 (1971) pp. 11-51.

[3] H. O. FATTORINI, *Un theorema de perturbatión para generatores de funciones coseno*, Rev. Un. Mat. Argentina, 25 (1971), 199-211.

[4a] ———, *Ordinary differential equations in linear topological spaces*. I, J. Differential Equations 5 (1968), pp. 72-105.

[4b] ———, *Ordinary differential equations in linear topological spaces* II, J. Differential Equations, 6 (1969), pp. 50-70.

[5] G. GRUBB, *Functional Calculus of Pseudo-Differential Boundary Problems*, Progress in Mathematics, Vol. 65, Birkhäuser, Boston, 1986.

[6] L. HÖRMANDER, *The Analysis of Linear Partial Differential Operators*, Vol. III, Springer-Verlag, Berlin, 1985.

[7] I. LASIECKA AND R. TRIGGIANI, *Feedback semigroups and cosine operators for boundary feedback parabolic and hyperbolic equations*, J. Differential Equations, 47 (1983), pp. 245-272.

[8] ———, *Stabilization and structural assignment of Dirichlet boundary feedback parabolic equations*, SIAM J. Control Optim., 21 (1983), pp. 766-802.

[9] ———, *Hyperbolic equations with Dirichlet boundary feedback via position vector*: *Regularity and almost periodic stabilization* I, Appl. Math. Optim., 8 (1981), pp. 1-37.

[10] J. L. LIONS AND E. MAGENES, *Non Homogeneous Boundary Value Problems and Applications*, Vols. I, II, Springer-Verlag, Berlin, New York, 1972.

[11] T. NAMBU, *Feedback stabilization for distributed parameter systems of parabolic type*, J. Differential Equations, 33 (1979), pp. 167-188.

[12] M. SOVA, *Cosine operator functions*, Razprawy Mat., 49 (1966), pp. 3-46.

[13] R. TRIGGIANI, *On Nambu's problem for diffusion processes*, J. Differential Equations, 33 (1979), pp. 189-200.

[14] ———, *Boundary feedback stabilizability of parabolic equations*, Appl. Math. Optim., 6 (1980), pp. 201-220.

[15] 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.

[16] W. M. WONHAM, *On pole assignment in multi-input controllable linear systems*, IEEE Trans. Automat. Control, 12 (1967), pp. 660-665.

[17] J. ZABCZYK, *On decomposition of generators*, SIAM J. Control Optim., 16 (1978), pp. 523-539.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 253-263, March 1991

VIABILITY PROBLEMS FOR NONAUTONOMOUS DIFFERENTIAL INCLUSIONS*

PETER TALLOS†

Abstract. In this paper the existence of viable solutions to nonautonomous differential inclusions is proved. The set-valued map on the right-hand side is assumed to be integrably bounded, measurable in t, and upper semicontinuous in x with nonempty convex, compact values. The cases of convex and nonconvex viability domains and also the time-dependent case are considered. An example for control systems is also given.

Key words. nonautonomous differential inclusions, measurable right-hand side, tangent cones, viable trajectories

AMS(MOS) subject classifications. 49A50, 34A60

1. Introduction. In this paper we investigate the existence of viable solutions to nonautonomous differential inclusions. More precisely, consider a finite-dimensional space X and let K be a nonempty closed subset of X. Let F be a set-valued map defined on $\mathbf{R}\times K$ with nonempty convex compact values in X. If x_0 is given in K, we look for a solution to the differential inclusion

$$x'(t)\in F(t, x(t)), \qquad x(0)=x_0, \tag{1}$$

which is *viable*, i.e.,

$$x(t)\in K \quad \text{for every } t\geqq 0. \tag{2}$$

Existence theorems for viable solutions to autonomous differential inclusions were proven by Haddad [10] (see also the book of Aubin and Cellina [2]). Moreover, the time-dependent viability theorem states that if for some $T>0$ the map F is bounded and upper semicontinuous on $[0, T]\times K$ and the tangential condition

$$F(t, x)\cap T_K(x)\neq\varnothing$$

holds true for every $(t, x)\in[0, T]\times K$, then for every x_0 in K there exists a viable solution defined on $[0, T]$ (see [2, Thm. 4.41]). Here $T_K(x)$ denotes the Bouligand contingent cone to K at x.

However, most of the classical existence theorems for differential inclusions (see, for instance, Filippov [7], Zaremba [16]) have been extended to the case of nonautonomous inclusions, when the set-valued map on the right-hand side is only measurable with respect to t (Himmelberg and Van Vleck [11], Olech [12], Plis [13]). Such differential inclusions are provided, for example, by "linearization" of an autonomous inclusion along a trajectory (see Frankowska [8], [9] for variational inclusions). It is a natural question to ask whether the time-dependent viability theorem remains true when only measurability in t is assumed for F and the boundedness assumption is replaced by integrably boundedness.

* Received by the editors December 7, 1987; accepted for publication (in revised form) May 29, 1990. This research was carried out in part while the author was visiting the International Institute for Applied Systems Analysis at Laxenburg, Austria.

† Department of Mathematics, Rutgers University, Camden, New Jersey 08102. Present address, Institute of Mathematics and Computer Sciences, University of Economics, Budapest, Hungary 1828.

In § 2 we give some preliminary facts, and in § 3 we provide an affirmative answer for convex viability domains. The proof is constructive and it appears as a refined version of Haddad's proof [10] (see also [2, Thm. 4.2.3]). The construction is based on the systematic use of the measurable selection theorem (see [2], [5]). A similar result was obtained by Deimling [6] by using a Scorza-Dragoni-type theorem of Rzeżuchowski [14] and reducing the problem to the autonomous case. In § 4 we deal with the nonconvex case and prove the existence of viable solutions under a stronger tangential condition. In particular, we use the Clarke tangent cone, since this cone is always convex, even if K is not. We define a sequence of upper semicontinuous approximations of F and use the time-dependent viability theorem. By passing to the limit, we get a viable solution of the original problem. Finally, in § 5 we consider time-dependent viability domains and an example for control systems is given.

2. Preliminaries. First we formulate the definition of integrably boundedness. The set-valued map F is said to be *integrably bounded*, if there exists a locally L^1 function l, i.e., a measurable function with $l \in L^1[0, T]$ for every $T > 0$ such that for almost every t in $\mathbf{R}$ and for every x in K

$$F(t, x) \subset l(t)(1+|x|)B$$

is valid, where B is the closed unit ball in X.

As is well known, if F is integrably bounded, by a change of t and a retraction in x, we may assume without loss of generality that for a $T > 0$

$$|v| \leqq 1 \tag{3}$$

for every $v \in F(t, x)$ and $(t, x) \in [0, T] \times K$.

The following two propositions will be used for proving the convergence of the sequence of approximate solutions. For the proofs we refer, for instance, to Olech [12].

LEMMA 1. *Suppose that F is integrably bounded and we are given two sequences $y_n \in L^1[0, T]$ and $x_n \in C[0, T]$ with*

$$\lim_{n\to\infty} d(y_n(t), F(t, x_n(t))) = 0 \quad a.e.,$$

where $y_n \to y$ weakly in L^1, $x_n \to x$ uniformly on $[0, T]$ and d denotes the distance function. Then

$$y(t) \in \bigcap_{m=1}^{\infty} \operatorname{cl} \operatorname{co} \bigcup_{n=m}^{\infty} F(t, x_n(t))$$

almost everywhere in $[0, T]$.

LEMMA 2. *Consider the set-valued map F and the sequence x_n in Lemma 1 and assume in addition that F is upper semicontinuous in x with convex compact values. Then*

$$\bigcap_{m=1}^{\infty} \operatorname{cl} \operatorname{co} \bigcup_{n=m}^{\infty} F(t, x_n(t)) = F(t, x(t))$$

for each t in $[0, T]$.

In the next sections we will use the concepts of the Bouligand contingent cone, the intermediate cone and the Clarke tangent cone. Recall that the *Bouligand contingent cone* to K at x is defined by

$$T_K(x) = \left\{ v \in X : \liminf_{h\to 0+} \frac{1}{h} d_K(x+hv) = 0 \right\},$$

where d_K denotes the distance function from the set K.

The *intermediate cone*, introduced by Ursescu [15], is defined by

$$I_K(x)=\left\{v\in X:\ \lim_{h\to 0+}\frac{1}{h}\,d_K(x+hv)=0\right\}.$$

The same concept was used by Frankowska in [8] for solving optimization problems for differential inclusions.

The *Clarke tangent cone* is defined by

$$C_K(x)=\left\{v\in X:\ \lim_{y\to x, h\to 0+}\frac{1}{h}\,d_K(y+hv)=0\right\}.$$

This cone is always convex.

Obviously, $C_K(x)\subset I_K(x)\subset T_K(x)$, and moreover, they coincide if K is convex. For details concerning various concepts of tangent cones and their properties we refer to Aubin and Ekeland [3] and Frankowska [8].

We will need the following simple proposition.

LEMMA 3. *Let $y\in K$ be fixed and define*

$$\Delta_h(y)=\sup_{v\in I_K(y)\cap B}\frac{1}{h}\,d_K(y+hv),$$

where B is the closed unit ball in X. Then

$$\lim_{h\to 0+}\Delta_h(y)=0.$$

Proof. Let $\varepsilon>0$ be given. Since $I_K(y)\cap B$ is compact we can find points $v_1,\cdots,v_m$ in $I_K(y)\cap B$ such that

$$I_K(y)\cap B\subset\bigcup_{i=1}^{m}\left(v_i+\frac{\varepsilon}{2}B\right).$$

On the other hand, for every $i=1,\cdots,m$ there exists a $\delta_i>0$ such that

$$\frac{1}{h}\,d_K(y+hv_i)<\frac{\varepsilon}{2}$$

if $h<\delta_i$. Put $\delta=\min\{\delta_1,\cdots,\delta_m\}$. Now take any v in $I_K\cap B$. Then there exists a j with $v\in v_j+(\varepsilon/2)B$. If $h<\delta$, by the Lipschitz continuity of d_K, we get

$$\frac{1}{h}\,d_K(y+hv)\leqq\frac{1}{h}\,d_K(y+hv_j)+|v-v_j|<\varepsilon,$$

consequently $\Delta_h(y)<\varepsilon$. □

3. Convex viability domains. In this section we prove the existence of viable solutions in the case of convex viability domains. For the construction we need the measurable selection theorem (see, for instance, Aubin and Cellina [2, p. 90]) and the convexity of K allows us to make use of the mean value theorem (see [2, p. 21]).

THEOREM 1. *Let K be closed, convex and suppose that F is integrably bounded, measurable in t, and upper semicontinuous in x with nonempty convex compact values. Assume that the tangential condition*

$$F(t,x)\cap T_K(x)\neq\varnothing \tag{4}$$

holds true for almost every t in $\mathbf{R}$ and for every x in K. Then for every $x_0\in K$ and for every $T>0$ there exists a viable solution to (1) *defined on* $[0, T]$.

Proof. Let $x_0 \in K$ be given and choose a $T > 0$ arbitrarily. Then, in view of (3), every trajectory for F through x_0 on $[0, T]$, if any exist, obviously lies in $x_0 + TB$, where B denotes the closed unit ball in X. On the other hand, $K \cap (x_0 + TB)$ is a compact subset of X and, since the contingent cone depends only on the local shape of K around x, for each $0 < \rho < T$ and $x \in K \cap (x_0 + \rho B)$, we clearly have that $T_K(x) = T_{K \cap (x_0 + TB)}(x)$. Hence, without loss of generality we may assume that K is compact.

Since K is convex, we have that $T_K(y) = I_K(y)$ for each $y \in K$. Fix an integer n; then, in view of Lemma 3, for every y in K there exists $0 < h_y < 1/n$ such that

$$\Delta_{h_y}(y) < \frac{1}{3nT}. \tag{5}$$

Since $T_K(y)$ is closed, we can find a measurable selection $f_y(t) \in F(t, y) \cap T_K(y)$ on $[0, T]$. Consider the sets

$$U(y) = \left\{ x \in X \colon d_K(x + h_y f_y(t)) < d_K(y + h_y f_y(t)) + \frac{h_y}{3nT} \text{ a.e.} \right\}.$$

By the Lipschitz continuity of d_K, $U(y)$ is open and $y \in U(y)$ for every $y \in K$, hence, there is a $0 < \delta_y < 1/n$ with $y + \delta_y B \subset U(y)$. Since K is compact, we can select finitely many points $y_1, \cdots, y_m$ in K such that

$$K \subset \bigcup_{i=1}^{m} (y_i + \delta_i B), \tag{6}$$

where $\delta_i = \delta_{y_i}$. Introduce the notation $h_i = h_{y_i}$, $f_i = f_{y_i}$ for $i = 1, \cdots, m$. Put $h_0(n) = \min\{h_1, \cdots, h_m\}$, then $0 < h_0(n) < 1/n$.

Now we construct the approximate solution x_n on the interval $[0, T]$. Set $t_0 = 0$ and $x_n(t_0) = x_0$. In view of (6), we can find an index $1 \leqq i \leqq m$ with $x_0 \in y_i + \delta_i B$. Thus, we get

$$d_K(x_0 + h_i f_i(t)) < d_K(y_i + h_i f_i(t)) + \frac{h_i}{3nT}.$$

Consider the following set-valued map on $[0, T]$:

$$Z_i(t) = K \cap \left\{ z \in X \colon z - x_0 - h_i f_i(t)| \leqq d_K(x_0 + h_i f_i(t)) + \frac{h_i}{3nT} \right\}.$$

Then Z_i is obviously measurable and admits nonempty closed values. Thus, by the measurable selection theorem, we can take a measurable selection z_i of Z_i with $z_i(t) \in K$ for almost every t in $[0, T]$.

Set $t_1 = h_i$, $v_0(t) = 1/h_i(z_i(t) - x_0)$ and define x_n on the interval $[t_0, t_1]$ by

$$x_n(t) = x_0 + \int_{t_0}^{t_1} v_0(s)\, ds.$$

Since K is convex, we obtain

$$x_n(t_1) = \frac{1}{h_i} \int_0^{h_i} z_i(t)\, dt \in K$$

by the mean value theorem.

The construction can be proceeded by induction. Suppose we have constructed x_n on the subinterval $[0, t_k]$ and $x_n(t_k) \in K$. In view of (6), we can find an index $1 \leqq j \leqq m$ such that $x_n(t_k) \in y_j + \delta_j B$, thus

$$d_K(x_n(t_k) + h_j f_j(t)) < d_K(y_j + h_j f_j(t)) + \frac{h_j}{3nT} \tag{7}$$

for almost every t. Introduce the following set-valued map on $[0, T]$:

$$Z_j(t) = K \cap \left\{ z \in X \colon |z - x_n(t_k) - h_j f_j(t)| \leqq d_K(x_n(t_k) + h_j f_j(t)) + \frac{h_j}{3nT} \right\}.$$

Then Z_j is clearly measurable with nonempty closed values. Take a measurable selection z_j of Z_j; then $z_j(t) \in K$ for almost every t in $[0, T]$.

Set $t_{k+1} = t_k + h_j$, $v_k(t) = 1/h_j(z_j(t) - x_n(t_k))$ and define x_n on the interval $[t_k, t_{k+1}]$ by

$$x_n(t) = x_n(t_k) + \int_{t_k}^{t} v_k(s)\, ds.$$

The convexity of K implies that

$$(8) \qquad x_n(t_{k+1}) = \frac{1}{h_j} \int_{t_k}^{t_k + h_j} z_j(t)\, dt \in K$$

by the mean value theorem. Therefore the construction can be continued.

Since $t_{k+1} - t_k \geqq h_0(n) > 0$ for each k, we reach the point T within finitely many steps. Hence, the approximate solution x_n can be defined on the whole interval $[0, T]$ and we obtain a partition $0 = t_0 < \cdots < t_p = T$ of $[0, T]$ for every n.

Now we prove the convergence of approximate solutions. We deduce from the above construction that each x_n is absolutely continuous and for almost every $t \in [t_k, t_{k+1}]$ we have

$$|x_n'(t) - f_j(t)| = \left| \frac{1}{h_j}(z_j(t) - x_n(t_k) - f_j(t)))\right|$$

$$\leqq \frac{1}{h_j} d_K(x_n(t_k) + h_j f_j(t)) + \frac{1}{3nT}$$

for a suitable index $1 \leqq j \leqq m$. In view of the definition of f_j, we get

$$(9) \qquad x_n'(t) \in F(t, y_j) + \left[\frac{1}{h_j} d_K(x_n(t_k) + h_j f_j(t)) + \frac{1}{3nT} \right] B.$$

For every $0 \leqq k \leqq p-1$ denote by i_k the corresponding index for which $x_n(t_k) \in y_{i_k} + \delta_{i_k} B$. Let us introduce the following functions on $[0, T]$:

$$e_k(t) = \begin{cases} 1 & \text{if } t \in [t_k, t_{k+1}], \\ 0 & \text{otherwise.} \end{cases}$$

Then (9) can be written in the following form:

$$(10) \qquad x_n'(t) \in F(t, x_n(t) + a_n(t)) + q_n(t) B$$

almost everywhere in $[0, T]$, where

$$a_n(t) = \sum_{k=0}^{p-1} y_{i_k} e_k(t) - x_n(t)$$

and

$$q_n(t) = \sum_{k=0}^{p-1} e_k(t) \frac{1}{h_{i_k}} d_K(x_n(t_k) + h_{i_k} f_{i_k}(t)) + \frac{1}{3nT}.$$

Making use of (5) and (7), we get

$$\|q_n\|_{L^1} = \sum_{k=0}^{p-1} \int_{t_k}^{t_{k+1}} \frac{1}{h_{i_k}} d_K(x_n(t_k) + h_{i_k} f_{i_k}(t))\, dt + \frac{1}{3n}$$

$$\leqq \sum_{k=0}^{p-1} \left[\Delta_{h_{i_k}}(y_{i_k}) + \frac{1}{3nT} \right] (t_{k+1} - t_k) + \frac{1}{3n} \leqq \frac{1}{n}.$$

Therefore, taking a subsequence, we may assume that $q_n \to 0$ almost everywhere in $[0, T]$.

On the other hand, for every n, x_n is absolutely continuous, $x_n(0) = x_0$, and for almost every $t \in [0, T]$

$$|x_n'(t)| \leqq 1 + \frac{2}{3nT} \tag{11}$$

by (9). Thus, by the Dunford-Pettis criterion we can select a subsequence, again denoted by x_n, which converges uniformly on $[0, T]$ to an absolutely continuous function x, moreover $x_n' \to x'$ weakly in $L^1[0, T]$. Furthermore, $a_n \to 0$ uniformly on $[0, T]$. Indeed, since $\delta_i < 1/n$ for each i, by (11) we have

$$|a_n(t)| = \left| \sum_{k=0}^{p-1} y_{i_k} e_k(t) - x_n(t) \right|$$

$$\leqq \sum_{k=0}^{p-1} |y_{i_k} - x_n(t_k)|\, e_k(t) + \left| \sum_{k=0}^{p-1} x_n(t_k)\, e_k(t) - x_n(t) \right|$$

$$\leqq \frac{1}{n} + \frac{1}{n}\left(1 + \frac{2}{3nT}\right).$$

Using Lemmas 1 and 2, we deduce from (10) that

$$x'(t) \in \bigcap_{m=1}^{\infty} \operatorname{cl\,co} \bigcup_{n=m}^{\infty} F(t, x_n(t) + a_n(t)) = F(t, x(t))$$

almost everywhere in $[0, T]$. Hence, in view of (8), we get that x is a viable solution to (1) on $[0, T]$. □

Remark. A similar result was proven by Deimling in [6]. The proof is based on the reduction of the problem to the autonomous case by using a Scorza-Dragoni type theorem of Rzeżuchowski [14].

4. Nonconvex viability domains. In this section we prove the existence of viable trajectories when K is closed, but not necessarily convex. The following method will be used. For every $h > 0$ we define an approximation F_h of the set-valued map F, which enjoys more regularity than F, then we apply Haddad's theorem for F_h. Passing to the limit $h \to 0$ we get a viable solution of the original inclusion. However, we need a stronger tangential condition in this case, namely, we use the Clarke tangent cone.

THEOREM 2. *Let K be nonempty closed and suppose that F is integrably bounded, measurable in t, and upper semicontinuous in x with nonempty convex compact values. We posit the tangential condition*

$$F(t, x) \cap C_K(x) \neq \varnothing \tag{12}$$

for almost every t in **R** *and every x in K. Then for every $x_0 \in K$ and $T > 0$ there exists a viable solution to* (1) *defined on* $[0, T]$.

Proof. Let $x_0 \in K$ and $T > 0$ be given. For every $h > 0$ define F_h by the Aumann-integral

$$F_h(t, x) = \frac{1}{h} \int_t^{t+h} F(s, x)\, ds,$$

where $(t, x) \in [0, T] \times K$. Then for all $h > 0$, for almost every $t \in [0, T]$ and for all $x \in K$ we have

$$F_h(t, x) \cap C_K(x) \neq \emptyset. \tag{13}$$

Indeed, take a measurable selection f of the measurable closed-valued map $s \to F(s, x) \cap C_K(x)$. Since $C_K(x)$ is convex, we get

$$\frac{1}{h} \int_t^{t+h} f(s)\, ds \in F_h(t, x) \cap C_K(x)$$

by the mean value theorem.

We show that F_h is upper semicontinuous on $[0, T] \times K$ for each fixed $h > 0$. Let $(t, x) \in [0, T] \times K$ and $\varepsilon > 0$ be given. Take an arbitrary $\gamma > 0$ and define

$$\eta(s, x, \gamma) = \sup \left\{ \sup_{u \in F(s,y)} d(u, F(s, x)) \colon |x - y| \leqq \gamma \right\}.$$

Then η is measurable in s (see [5, Thm. III.9, Lemma III.39]) and by (3), $\eta(s, x, \gamma) \leqq 2$ for almost every $s \in [0, T]$ and for every $x \in K$ and $\gamma > 0$. Moreover, $\eta(s, x, \gamma) \to 0$ if $\gamma \to 0$, by the upper semicontinuity of F. Hence, by Lebesgue's dominated convergence theorem

$$\lim_{\gamma \to 0} \frac{1}{h} \int_t^{t+h} \eta(s, x, \gamma)\, ds = 0.$$

Choose a $0 < \delta < \varepsilon h/4$ such that

$$\frac{1}{h} \int_t^{t+h} \eta(s, x, \delta)\, ds < \frac{\varepsilon}{2} \tag{14}$$

is valid.

Now take an $(r, y) \in [0, T] \times K$ such that $|t - r| < \delta$, $|x - y| < \delta$ and pick $w \in F_h(r, y)$. In view of the definition of F_h, we can find a measurable selection $g(s) \in F(s, y)$ with

$$w = \frac{1}{h} \int_r^{r+h} g(s)\, ds.$$

Consider the following set-valued map:

$$\Phi(s) = F(s, x) \cap \{u \in X \colon |u - g(s)| \leqq \eta(s, x, \delta)\}.$$

Then Φ is obviously measurable with nonempty closed values. Let f be a measurable selection of Φ and set

$$v = \frac{1}{h} \int_t^{t+h} f(s)\, ds.$$

Then $v \in F_h(t, x)$, and by (14) we have

$$\begin{aligned}|v-w| &= \frac{1}{h}\left|\int_t^{t+h} f(s)\,ds - \int_r^{r+h} g(s)\,ds\right| \\ &\leqq \frac{1}{h}\left|\int_t^r |g(s)|\,ds\right| + \frac{1}{h}\int_t^{t+h} |f(s)-g(s)|\,ds + \frac{1}{h}\left|\int_{t+h}^{r+h} |g(s)|\,ds\right| \\ &\leqq \frac{\varepsilon}{4} + \frac{1}{h}\int_t^{t+h} \eta(s, x, \delta)\,ds + \frac{\varepsilon}{4}.\end{aligned}$$

Thus, F_h is upper semicontinuous on $[0, T] \times K$.

Since for every $h > 0$, F_h is a bounded map with nonempty convex compact values, furthermore the tangential condition (13) is valid, we can apply the time-dependent viability theorem (see [1, Thm. 4.4.1]). Hence, for every $x_0 \in K$ and $T > 0$ there exists a viable trajectory x_h for F_h through x_0 defined on $[0, T]$. This means that

$$\begin{aligned}&x_h'(t) \in F_h(t, x_h(t)) \quad \text{a.e. in } [0, T], \\ &x_h(0) = x_0, \\ &x_h(t) \in K \quad \text{for every } t \text{ in } [0, T].\end{aligned}$$

By the boundedness assumption (3), there exists a subsequence denoted by $x_n = x_{h_n}$ that converges uniformly on $[0, T]$ to an absolutely continuous function x with $h_n \to 0$ and $x_n' \to x'$ weakly in $L^1[0, T]$.

Since the set-valued map $t \to F(t, x(t))$ can be assumed to be measurable (see [14]), it follows that the map $t \to \eta(t, x(t), \gamma)$ is also measurable for every $\gamma > 0$ (see [5, Chap. 3]). On the other hand, $\eta(t, x(t), \gamma) \to 0$ almost everywhere, if $\gamma \to 0+$.

Let δ_n be any sequence converging decreasingly to zero. We show that

$$\lim \frac{1}{h_n}\int_t^{t+h_n} \eta(s, x(s), \delta_n)\,ds = 0 \tag{15}$$

almost everywhere in $[0, T]$. Indeed, it is clear that there is a set $H \subset [0, T]$ of measure T such that for every $t \in H$

$$\lim_{h\to 0+} \frac{1}{h}\int_t^{t+h} \eta(s, x(s), \delta_n)\,ds = \eta(t, x(t), \delta_n)$$

for each n. In fact, H is a countable intersection of sets of measure T. Take a point $t \in H$ and let $\varepsilon > 0$ be given. Then we can find an n_0 with $\eta(t, x(t), \delta_{n_0}) < \varepsilon/2$, and an n_1 such that

$$\frac{1}{h_n}\int_t^{t+h_n} \eta(s, x(s), \delta_{n_0})\,ds < \eta(t, x(t), \delta_{n_0}) + \frac{\varepsilon}{2}$$

for $n \geqq n_1$. Hence, if $n \geqq \max\{n_0, n_1\}$ we have

$$\begin{aligned}\frac{1}{h_n}\int_t^{t+h_n} \eta(s, x(s), \delta_n)\,ds &\leqq \frac{1}{h_n}\int_t^{t+h_n} \eta(s, x(s), \delta_{n_0})\,ds \\ &< \eta(t, x(t), \delta_{n_0}) + \frac{\varepsilon}{2} < \varepsilon,\end{aligned}$$

because η is increasing with respect to δ.

Pick a point $s \in [t, t+h_n]$ and introduce the notation

$$|x_n(t) - x(s)| \leqq |x_n(t) - x(t)| + h_n = \delta_n.$$

We may assume (by taking a subsequence if necessary) that δ_n converges decreasingly to zero. Moreover, we have

$$(16)\qquad \begin{aligned} x_n'(t) &\in \frac{1}{h_n}\int_t^{t+h_n} F(s, x_n(t))\,ds \\ &\subset \frac{1}{h_n}\int_t^{t+h_n} F(s, x(s))\,ds + \frac{1}{h_n}\int_t^{t+h_n} \eta(s, x(s), \delta_n)\,ds\,B. \end{aligned}$$

Making use of (15), (16) and the Lebesgue-point equality

$$(17)\qquad \lim_{n\to\infty} \frac{1}{h_n}\int_t^{t+h_n} F(s, x(s))\,ds = F(t, x(t))$$

(see [4]), we get

$$(18)\qquad d(x_n'(t), F(t, x(t))) \to 0$$

almost everywhere in $[0, T]$. Thus, by Lemmas 1 and 2, relation (18) and the obvious $x(t) \in K$ show that x is a viable solution to (1) defined on $[0, T]$. □

5. The time-dependent case. If the viability domain K depends on t we can prove the existence of viable trajectories by standard arguments. We express the tangential condition in terms of the graphical derivative of the set-valued map K (see Aubin and Ekeland [3]).

Let K be a set-valued map defined on the interval $[0, T]$ with nonempty values in X. Recall that the *graphical derivative* of K at $(t, x) \in \operatorname{graph} K$ is defined as the set-valued map $DK(t, x)$, whose graph is $C_{\operatorname{graph} K}(t, x)$. More precisely $v \in DK(t, x)(u)$ if and only if $(u, v) \in C_{\operatorname{graph} K}(t, x)$. Several properties of this derivative, including chain rule and inverse function theorem, can be found in [3] and [1].

Now consider a set-valued map F defined on graph K with nonempty convex compact values in X, and let x_0 be a given point in $K(0)$. A solution to the differential inclusion problem

$$(19)\qquad x'(t) \in F(t, x(t)), \qquad x(0) = x_0$$

is said to be *viable* if

$$(20)\qquad x(t) \in K(t)$$

for every $t \in [0, T]$.

THEOREM 3. *Let the graph of K be closed and suppose that F is integrably bounded, measurable in t, and upper semicontinuous in x with nonempty convex compact values. Assume that the tangential condition*

$$(21)\qquad F(t, x) \cap DK(t, x)(1) \neq \emptyset$$

holds true for almost every $t \in [0, T]$ *and every* x *in* $K(t)$. *Then for every* x_0 *in* $K(0)$ *there exists a viable solution to* (19) *defined on* $[0, T]$.

Proof. By using the usual transformation of the state space with a trivial modification, it can easily be seen that all assumptions of Theorem 2 are satisfied. Moreover, the tangential condition (21) reduces to (12), hence, the statement of the theorem follows from Theorem 2. □

Remark. The concept of the graphical derivative of a set-valued map clearly depends on what tangent cone is chosen. Therefore, if the graph of K is convex and closed, the above theorem can be reformulated in terms of the *contingent derivative* of K, i.e., by replacing the Clarke tangent cone with the Bouligand contingent cone.

As an illustration of the above theorem consider the following example. Let X and Y be finite-dimensional spaces. Let K be a set-valued map defined on the interval $[0, T]$ with nonempty values in X, and let $x_0 \in K(0)$ be given. Consider the control differential equation

$$x'(t) = f(t, x(t), u(t)), \qquad x(0) = x_0, \tag{22}$$

where the control set $U \subset Y$ is compact, the controls are the measurable functions u with $u(t) \in U$ almost everywhere in $[0, T]$, and f: graph $K \times U \to X$ is integrably bounded, measurable in t, and continuous in (x, u).

The problem can be expressed in the following way. Does there exist a control u such that the control system (22) has a viable trajectory, i.e., a solution x such that $x(t) \in K(t)$ for every $t \in [0, T]$? We solve this problem by rewriting the differential equation (22) as a differential inclusion and applying Theorem 3. As is well known, by Filippov's implicit function lemma (see [5, p. 85]), the differential equation (22) and the corresponding differential inclusion have the same set of trajectories.

We will use the following assumption:

$$f(t, x, U) \subset X \text{ is convex} \tag{23}$$

for almost every $t \in [0, T]$ and every $x \in K(t)$. This condition is satisfied, for instance, if f is affine with respect to u and U is convex.

Let us introduce the feedback map C defined by

$$C(t, x) = \{u \in U : f(t, x, u) \in DK(t, x)(1)\}$$

for every $(t, x) \in$ graph K.

THEOREM 4. *Let the graph of K be closed and suppose that the feedback map*

$$C(t, x) \neq \emptyset \tag{24}$$

for every $(t, x) \in$ graph K. Then, under the above assumptions, for every $x_0 \in K(0)$ there exist a control u and a viable trajectory x of (22) *that are related by*

$$u(t) \in C(t, x(t)) \tag{25}$$

almost everywhere in $[0, T]$.

Proof. Introduce the following set-valued map on graph K:

$$F(t, x) = \{f(t, x, u) : u \in U\}.$$

Then F clearly satisfies all assumptions of Theorem 3. Moreover, by (24), the tangential condition (21) is also fulfilled. Thus, there exists a viable trajectory x for F through x_0. By Filippov's implicit function lemma, we can find a control u such that

$$x'(t) = f(t, x(t), u(t))$$

almost everywhere in $[0, T]$. This means that x is a solution to (22). Since obviously $x'(t) \in DK(t, x(t))(1)$, we have that (25) is also satisfied. □

A detailed discussion of the above problem for autonomous equations and time-independent viability domains can be found in [1] and [2].

Acknowledgments. The author is greatly indebted to J.-P. Aubin and H. Frankowska for helpful discussions. The author also thanks the referees for critical comments to the earlier versions of the paper.

REFERENCES

[1] J.-P. AUBIN, *Viability Theory*, to appear.

[2] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[3] J.-P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, John Wiley, New York, 1984.

[4] T. F. BRIDGLAND, JR., *Trajectory integrals of set valued functions*, Pacific J. Math., 33 (1970), pp. 43-68.

[5] C. CASTAING AND M. VALADIER, *Convex Analysis and Measurable Multifunctions*, Lecture Notes in Mathematics, Vol. 580, Springer-Verlag, Berlin, 1977.

[6] K. DEIMLING, *Extremal solutions of multivalued differential equations* II, Results in Math., 15 (1989), pp. 197-201.

[7] A. F. FILIPPOV, *On the existence of solutions to multivalued differential equations*, Mat. Zametki, 10 (1971), pp. 307-313. (In Russian.)

[8] H. FRANKOWSKA, *The maximum principle for an optimal solution to a differential inclusion with end points constraints*, SIAM J. Control Optim., 25 (1987), pp. 145-157.

[9] ———, *Local controllability and infinitesimal generators of semigroups of set-valued maps*, SIAM J. Control Optim., 25 (1987), pp. 412-432.

[10] G. HADDAD, *Monotone trajectories of differential inclusions and functional differential inclusions with memory*, Israel J. Math., 39 (1981), pp. 83-100.

[11] C. J. HIMMELBERG AND F. S. VAN VLECK, *Existence of solutions for generalized differential equations with unbounded right-hand side*, J. Differential Equations, 61 (1986), pp. 295-320.

[12] C. OLECH, *Existence of solutions of nonconvex orientor fields*, Boll. Un. Mat. Ital., 11 (1975), pp. 189-197.

[13] A. PLIS, *Measurable orientor fields*, Bull. Acad. Polon. Sci. Sér. Math., 13 (1965), pp. 565-569.

[14] T. RZEŻUCHOWSKI, *Scorza-Dragoni type theorem for upper semicontinuous multivalued functions*, Bull. Acad. Polon. Sci. Sér. Math., 28 (1980), pp. 61-66.

[15] C. URESCU, *Tangent set's calculus and necessary conditions for extremality*, SIAM J. Control Optim., 20 (1982), pp. 563-574.

[16] S. C. ZAREMBA, *Sur les équations au paratingent*, Bull. Sci. Math. (2), 60 (1936), pp. 139-160.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 264-282, March 1991

CONTINUOUS-TIME STOCHASTIC ADAPTIVE CONTROL*

M. GEVERS†, G. C. GOODWIN‡, AND V. WERTZ†

Abstract. This paper establishes global boundedness for a continuous-time stochastic adaptive control algorithm. It is shown that, with probability one, the system inputs and outputs satisfy a sample mean square boundedness property. The algorithm and method of analysis are not directly analogous to the discrete-time case, since special features are necessary to handle the continuous-time problem.

Key words. global boundedness, stochastic adaptive control

1. Introduction. It is now well known that global boundedness can be established for discrete-time stochastic adaptive control algorithms—see, for example, [1]-[3].

However, to date there has been no corresponding result for continuous-time systems. Preliminary work in this direction appears in [4]-[8]. However, the available adaptive control results rely upon unproven and data dependent conjectures, e.g., the normalized regression vector, $(\varphi^T\varphi/r)$, is assumed to be uniformly bounded. Inspection of the details of the algorithms indicates that this is most likely not true. A further restriction in the work reported in [5] and [6] is that the relative degree of the system is taken to be zero, i.e., there is a direct feedthrough term between input and output. This leads to questions about the validity of the resulting control law as the closed-loop system contains an algebraic loop.

This paper presents a new algorithm for continuous-time stochastic adaptive control in the ideal case (no unmodelled dynamics) together with a proof of global boundedness. The main feature of the algorithm is a new normalization technique that guarantees that the normalized regression vector $(\varphi^T\varphi/r)$ is bounded. Also, the proof of boundedness of the system states has several novel features including a special technique for handling strictly proper systems.

The paper was inspired by an earlier paper [9] which explored the link between discrete- and continuous-time *deterministic* adaptive control theory. The current paper does this for the stochastic case.

Novel aspects of our analysis are the following: We first establish existence and uniqueness for the solutions of the nonlinear stochastic differential equation for the parameter estimation algorithm. We then establish properties of the parameter estimator that hold irrespective of the control law. Finally, we establish a continuous-time key technical lemma that is analogous to that given for continuous-time deterministic systems in [9].

These preliminary results can be combined with a wide class of certainty equivalence control laws to establish global boundedness of all internal variables of the resultant algorithm. We illustrate by using an adaptive pole assignment algorithm as in [10] and [11] for deterministic continuous time systems.

The results are believed to be of importance, in their own right, insofar as they formally establish the boundedness of all system states and parameters for a continuous-time stochastic adaptive control algorithm. However, they also give further insight into

* Received by the editors December 30, 1987; accepted for publication (in revised form) February 8, 1990.

† Laboratoire d'Automatique, Louvain University, Bâtiment Maxwell, B-1348 Louvain-La-Neuve, Belgium.

‡ Department of Electrical and Computer Engineering, University of Newcastle, New South Wales, Australia.

the discrete-time theory, since they show what happens in the case of rapid sampling as shown in [18]. We have also recently [22] built on the results established here to establish global boundedness for a stochastic model reference adaptive control algorithm which is the continuous-time counterpart of the discrete-time stochastic adaptive minimum variance control results given in [1].

2. The model. In previous papers dealing with continuous-time stochastic adaptive control, an integral operator model has been used (see for example, [4]-[8]). Here, however, we adopt the more conventional continuous-time state space innovations representation, i.e.,

$$dx_t = Ax_t\,dt + Bu_t\,dt + K\,d\omega_t \tag{2.1}$$

$$dy_t = Cx_t\,dt + d\omega_t \tag{2.2}$$

where ω_t is a Wiener process with incremental covariance $\sigma^2\,dt$, x_t is a state vector of dimension n, u_t is a scalar control input, and dy_t is the scalar output of the system. The matrices A, B, K, C contain unknown but fixed parameters. We will denote by $\mathcal{F}_t$ the increasing σ-fields generated by $\{\omega_s, 0 \leqq s \leqq t\}$ and the unknown initial conditions x_0, and we will assume that $\|x_0\|^2$ is bounded. This model is the basis of stochastic control in the nonadaptive literature [12], [15]. It has also been proposed for the adaptive case [7]. Notice that y_t is the integral of the output.

Without loss of generality, we assume that the above model is in observer form where

$$A = \begin{bmatrix} -a_{n-1} & 1 & & \\ -a_{n-2} & & \ddots & \\ \vdots & & & 1 \\ -a_0 & & & 0 \end{bmatrix}; \quad B = \begin{bmatrix} b_{n-1} \\ \vdots \\ b_0 \end{bmatrix}; \quad K = \begin{bmatrix} k_{n-1} \\ \vdots \\ k_0 \end{bmatrix} \tag{2.3}$$

$$C = [1 \quad 0 \quad 0]. \tag{2.4}$$

For the purpose of adaptive control it is convenient to reexpress this model in fractional form [9]. We therefore reparametrize the system as follows. Let

$$E(\rho) = \rho^n + e_{n-1}\rho^{n-1} + \cdots + e_0 \tag{2.5}$$

$$G^T = [g_{n-1}, \cdots, g_0] \quad \text{with } g_i = e_i - a_i; \qquad i = 0, \cdots, n-1 \tag{2.6}$$

and

$$E = \begin{bmatrix} -e_{n-1} & 1 & & \\ \vdots & & \ddots & \\ \vdots & & & 1 \\ -e_0 & & 0 & 0 \end{bmatrix} \tag{2.7}$$

where the coefficients are arbitrary subject to $E(\rho)$ having all its zeros in the open left half plane and $K \neq G$.

Then, (2.1), (2.2) can be rewritten as

$$dx_t = Ex_t\,dt + Bu_t\,dt + (K - G)\,d\omega_t + G\,dy_t \tag{2.8}$$

$$dy_t = Cx_t\,dt + d\omega_t. \tag{2.9}$$

Using superposition, (2.8), (2.9) can also be expressed as

$$d\phi_t^1 = E\phi_t^1\,dt + G\,dt_t \tag{2.10}$$

$$d\phi_t^2 = E\phi_t^2\,dt + Bu_t\,dt \tag{2.11}$$

$$d\phi_t^3 = E\phi_t^3\,dt + (K-G)\,d\omega_t \tag{2.12}$$

$$dy_t = C[\phi_t^1\,dt + \phi_t^2\,dt + \phi_t^3\,dt] + d\omega_t. \tag{2.13}$$

Since y_t is a scalar, we have that $C(sI-E)^{-1}B = B^T(sI-E^T)^{-1}C^T$, etc. Hence (2.10) to (2.13) can be rewritten as

$$d\phi_t^y = E^T\phi_t^y\,dt + C^T\,dy_t \tag{2.14}$$

$$d\phi_t^u = E^T\phi_t^u\,dt + C^Tu_t\,dt \tag{2.15}$$

$$d\phi_t^\omega = E^T\phi_t^\omega\,dt + C^T\,d\omega_t \tag{2.16}$$

$$dy_t = [G^T\phi_t^y + B^T\phi_t^u + (K-G)^T\phi_t^\omega]\,dt + d\omega_t. \tag{2.17}$$

In (2.14)-(2.16) we choose $\phi_0^y = 0$, $\phi_0^u = 0$, and ϕ_0^ω such that

$$(K-G)^T\phi_0^\omega = Cx_0. \tag{2.18}$$

Equation (2.17) is in the form of a linear regression, i.e.,

$$dy_t = \phi_t^T\theta\,dt + d\omega_t \tag{2.19}$$

where

$$\phi_t^T = [(\phi_t^y)^T, (\phi_t^u)^T, (\phi_t^\omega)^T] \tag{2.20}$$

$$\theta^T = [G^T, B^T, (K-G)^T] = [G^T, B^T, F^T]. \tag{2.21}$$

Note that ϕ_t does not depend on the unknown system parameters since E is known. Thus, (2.14)-(2.16) simply represent a state space form of the usual regression vector as in [2] and [9]. To make the comparison with the regression vector formulations more complete, we might note that with $\rho \triangleq (d/dt)$ we have, with some abuse of notation and ignoring initial conditions,

$$(\phi_t^u)^T = \left[\frac{\rho^{n-1}}{E(\rho)}u, \cdots, \frac{1}{E(\rho)}u\right] \tag{2.22}$$

and similarly for ϕ_t^y, ϕ_t^ω. In the following, however, we will use the rigorous state space formulation as outlined earlier.

3. Pseudolinear regression estimation algorithm. The model (2.14)-(2.17) is not quite in a form that is suitable for parameter estimation. This is because the component ϕ_t^ω depends on the unmeasured noise source ω_t. Thus, as in the discrete case [1]-[3], [13], we define the predicted output by a pseudoregression in which $d\omega_t$ is replaced by the prediction error. Thus we define

$$d\hat{y}_t = \psi_t^T\hat{\theta}_t\,dt \tag{3.1}$$

where $\hat{\theta}_t$ is some bounded $\mathcal{F}_t$-measurable function. Later in this section we will define $\hat{\theta}_t$ as an estimate of θ using a stochastic differential equation driven by the data y_t, u_t. Also, in (3.1) we have

$$\psi_t^T = [(\psi_t^y)^T, (\psi_t^u)^T, (\psi_t^e)^T] \tag{3.2}$$

$$d\psi_t^y = E^T\psi_t^y\,dt + C^T\,dy_t \tag{3.3}$$

$$d\psi_t^u = E^T\psi_t^u\,dt + C^Tu_t\,dt \tag{3.4}$$

$$d\psi_t^e = E^T\psi_t^e\,dt + C^T\,de_t \tag{3.5}$$

$$de_t = dy_t - d\hat{y}_t = dy_t - \psi_t^T\hat{\theta}_t\,dt. \tag{3.6}$$

It is assumed that $\psi_0 = 0$ and that $\hat{\theta}_0$ is $\mathscr{F}_0$-measurable. With this choice of initial conditions, $\psi_t^y = \phi_t^y$ and $\psi_t^u = \phi_t^u$ for $t \geqq 0$.

The effect of using pseudoregressions can be further clarified as follows. We define

$$\eta_t = \phi_t^T \theta - \psi_t^T \hat{\theta}_t. \tag{3.7}$$

We then obtain

$$de_t - d\omega_t = \eta_t \, dt. \tag{3.8}$$

Note that η_t is the "deterministic part" of the prediction error. Now let

$$\gamma_t \triangleq \psi_t^e - \phi_t^\omega. \tag{3.9}$$

Then from (3.5), (3.7), (2.16) we have

$$d\gamma_t = E^T \gamma_t \, dt + C^T \eta_t \, dt, \quad \text{with } \gamma_0 = -\phi_0^\omega. \tag{3.10}$$

Noting that

$$\psi_0^y = \phi_0^y = 0 \quad \text{and} \quad \psi_0^u = \phi_0^u = 0, \tag{3.11}$$

then

$$(K - G)^T \gamma_t = -(\phi_t - \psi_t)^T \theta. \tag{3.12}$$

In particular,

$$(K - G)^T \gamma_0 = -Cx_0. \tag{3.13}$$

Note that γ_0 cannot be made zero if x_0 is unknown, but we will assume that $\|\gamma_0\|^2$ is bounded; this is consistent with our assumption on x_0 and (3.13).

Denoting

$$\tilde{\theta}_t \triangleq \hat{\theta}_t - \theta \tag{3.14}$$

we then have from (3.12), (3.10) that

$$\eta_t = -(K - G)^T \gamma_t - \psi_t^T \tilde{\theta}_t \tag{3.15a}$$

where

$$d\gamma_t = (A - KC)^T \gamma \, dt + C^T(-\psi_t^T \tilde{\theta}_t) \, dt. \tag{3.15b}$$

We thus see that η_t is related to $-\psi_t^T \tilde{\theta}_t$ by the following transfer function equation

$$D(\rho)\eta_t = E(\rho)[-\psi_t^T \tilde{\theta}_t] \tag{3.16}$$

where $D(\rho)$ is the characteristic polynomial of the optimal Kalman filter for the system, i.e.,

$$D(\rho) = \rho^n + (k_{n-1} + a_{n-1})\rho^{n-1} + \cdots (k_0 + a_0) = \det(\rho I - A + KC) \tag{3.17}$$

and where $E(\rho)$ is as in (2.5).

As is standard in pseudoregression algorithms [3], we require that sufficient prior knowledge is available to choose the observer polynomial $E(\rho)$ so as to satisfy the following assumption.

Assumption 1. $D(\rho)$ is strictly Hurwitz and the filter $E(\rho)$ is chosen such that

(1) $\operatorname{Re} \sigma_i(E(\rho)) \leqq -\alpha < 0$, $i = 1, \cdots, n$ where $\sigma_i(E)$ are the roots of E.

(2) D/E is input strictly passive, i.e., there exists $\varepsilon > 0$ and $K > 0$ such that

$$\forall T > 0, \int_0^T y_\tau u_\tau \, d\tau \geqq \varepsilon \int_0^T u_\tau^2 \, d\tau - K\|\gamma_0\|^2 \tag{3.18}$$

where y_t is the output of the filter D/E driven by u_t.

We will later need the following technical result on input strictly passive systems.

LEMMA 3.1. *Let $H(s)$ be input strictly passive with impulse response h, and let $y=h_* u$. Let $r_t>0$ be monotonically nondecreasing. Then there exists $\varepsilon>0$ such that for all $T>0$ and all u_t:*

$$\int_0^T \frac{y_t u_t}{r_t}\,dt \geqq \varepsilon \int_0^T \frac{u_t^2}{r_t}\,dt - \frac{K\|\gamma_0\|^2}{r_0}. \tag{3.19}$$

Proof. The proof follows immediately from (3.18) on using the result in Appendix A. □

Motivated by the discrete-time algorithm given in [1], we next define the estimation algorithm as

$$d\hat{\theta}_t = \frac{\psi_t}{r_t}(dy_t - \psi_t^T\hat{\theta}_t\,dt) \tag{3.20}$$

where

$$r_t \triangleq \sup_{0\leqq\tau\leqq t} \psi_\tau^T\psi_\tau + \int_0^t \psi_\tau^T\psi_\tau\,d\tau + c_0; \qquad c_0>0 \tag{3.21}$$

and c_0 is any positive deterministic number. Note that $\hat{\theta}_t$ is obtained by an Ito integral which makes sense locally since (ψ_t/r_t) is $\mathcal{F}_t$-measurable and continuous.

The definition of r_t given above is not quite the analogue of that used in discrete time. We will discuss the reason for the difference later.

Noting that $d\theta=0$ and using the definition of e_t and η_t given in (3.6), (3.8), we observe from (3.14) and (3.20) that $\tilde{\theta}_t$ is the solution of the following stochastic differential equation

$$d\tilde{\theta}_t = \frac{\psi_t}{r_t}\eta_t\,dt + \frac{\psi_t}{r_t}\,d\omega_t. \tag{3.22}$$

In the following we will require that $\tilde{\theta}_t$ be bounded. We guarantee this by introducing a projection scheme as described below. We first introduce the following assumption.

Assumption 2. There exists a known parameter value θ_c and a positive number R_1 such that the true value θ lies inside $\mathcal{C}_1$ where

$$\mathcal{C}_1 = \{\theta: \|\theta-\theta_c\| \leqq R_1\}.$$

Let R_2 be another positive number larger than R_1. We then modify the parameter estimator to ensure that $\|\hat{\theta}_t-\theta_c\|<R_2$ for all t. We do this by using the following projection scheme.

Parameter estimator with projection. Let τ be a time for which the solution of (3.20) is such that $\|\hat{\theta}_\tau-\theta_c\|=R_2$. Denote the corresponding value of $\hat{\theta}_\tau$ by $\hat{\theta}_{\tau_-}$. At time τ, the estimate $\hat{\theta}_\tau$ is then defined as

$$\hat{\theta}_\tau \triangleq \theta_c + \frac{R_1}{R_2}(\dot{\hat{\theta}}_{\tau_-} - \theta_c). \tag{3.23}$$

For $t\geqq\tau$, (3.20) is then integrated with initial condition $\hat{\theta}_\tau$ defined by (3.23). This makes $\hat{\theta}_t$ right continuous at the projection times.

4. A general class of feedback control laws. We consider a general class of control laws in state feedback form:

$$u_t = -[l_{n-1},\cdots,l_0]\psi_t^u - [p_{n-1},\cdots,p_0]\psi_t^y + y^*. \tag{4.1}$$

This is equivalent to the feedback law

$$Q(\rho)u_t = -P(\rho)y_t + E(\rho)y_t^* \tag{4.2}$$

where

$$Q(\rho) = E(\rho) + L(\rho);\ L(\rho) = l_{n-1}\rho^{n-1} + \cdots + l_0;\ P(\rho) = p_{n-1}\rho^{n-1} + \cdots + p_0. \tag{4.3}$$

Note that the control law transfer function is $-P/Q$, which is strictly proper. Also, y_t^* denotes a bounded reference signal. For the moment, we make no assumptions about L, P stabilizing the system or whether the control law depends on the estimated parameters $\hat{\theta}$. To allow for the latter possibility, we express the control law as:

$$u_t = -[\hat{l}_{n-1}, \cdots, \hat{l}_0]\psi_t^u - [\hat{p}_{n-1}, \cdots, \hat{p}_0]\psi_t^y + y_t^*. \tag{4.4}$$

For the moment, the only restriction we place on the above general control law is that $\hat{l}_{n-1}, \cdots, \hat{l}_0, \hat{p}_{n-1}, \cdots, \hat{p}_0$ by Lipschitz functions of $\hat{\theta}$.

From the model (2.1), (2.2), the general controller (4.4), the definition of ψ_t ((3.2)-(3.5)), and the definition of the errors ((3.6)-(3.8)), we can write:

$$d\psi_t = A_t\psi_t\,dt + B_1(\eta_t\,dt + d\omega_t) + B_2 y_t^*\,dt \tag{4.5}$$

where

$$A_t = \left[\begin{array}{ccc|ccc|ccc}
-\hat{a}_{n-1} & \cdots & -\hat{a}_0 & \hat{b}_{n-1} & \cdots & \hat{b}_0 & \hat{f}_{n-1} & \cdots & \hat{f}_0 \\
 & & 0 & & & & & & \\
 & I_{n-1} & \vdots & & & & & & \\
\cdots & & 0 & \cdots & & \cdots & \cdots & & \cdots \\
\hline
-\hat{p}_{n-1} & \cdots & -\hat{p}_0 & -\hat{q}_{n-1} & \cdots & -\hat{q}_0 & 0 & \cdots & 0 \\
 & & & & & 0 & & & \\
 & & & & I_{n-1} & \vdots & & & \\
\cdots & & \cdots & \cdots & & 0 & \cdots & & \cdots \\
\hline
0 & \cdots & 0 & 0 & \cdots & 0 & -e_{n-1} & \cdots & -e_0 \\
 & & & & & & & & 0 \\
 & & & & & & & I_{n-1} & \vdots \\
\cdots & & \cdots & \cdots & & \cdots & \cdots & & 0
\end{array}\right] \tag{4.6a}$$

$$B_1^T = [10 \cdots 010 \cdots 0] \tag{4.6b}$$

$$B_2^T = [0 \cdots 010 \cdots 0] \tag{4.6c}$$

where B_1^T has 1's in the first and $(2n+1)$st positions and B_2^T has 1 in the $(n+1)$st position.

A key point about (4.5) is that A_t is a Lipschitz function of $\tilde{\theta}_t$ provided the general control law (4.2) is chosen in which $\hat{l}_{n-1}, \cdots, \hat{l}_0, \hat{p}_{n-1}, \cdots, \hat{p}_0$ are Lipschitz in $\tilde{\theta}$.

5. General properties of the estimation algorithm with feedback. We first address the question of existence and uniqueness [14] of the solution of the full set of equations describing the system and estimation algorithm. Combining (3.14), (3.22), and (4.5), the full set of equations is

$$d\psi_t = \left[A_t(\tilde{\theta}_t)\psi_t - B_1\psi_t^T\tilde{\theta}_t - B_1\int_0^t h_{t-\tau}\psi_\tau^T\tilde{\theta}_\tau\,d\tau\right]dt + B_1\,d\omega_t + B_2 y_t^*\,dt \tag{5.1a}$$

$$d\tilde{\theta}_t = \left[-\frac{\psi_t\psi_t^T}{r_t(\psi.)}\tilde{\theta}_t - \int_0^t h_{t-\tau}\frac{\psi_t\psi_\tau^T}{r_t(\psi.)}\tilde{\theta}_\tau\,d\tau\right]dt + \frac{\psi_t}{r_t(\psi.)}\,d\omega_t \tag{5.1b}$$

where $A_t(\tilde{\theta}_t) \equiv A_t(\hat{\theta}_t)$ and h_t is the impulse response of the strictly proper part of the transfer function E/D in (3.16).

We then have the following result.

LEMMA 5.1. *The composite set of* (5.1) *has a unique solution with continuous sample paths almost surely up to the random time T of the first explosion.* (*That is, T is the first time that either a component of ψ or $\tilde{\theta}$ becomes infinite or $T=\infty$.*)

Proof. We first note from (4.6a) that $A_t(\tilde{\theta}_t)$ is Lipschitz in $\tilde{\theta}_t$ due to the assumed form of the dependence of $\hat{l}_{n-1}, \cdots, \hat{l}_0, \hat{p}_{n-1}, \cdots, \hat{p}_0$ on $\hat{\theta}$. This implies that the coefficient vectors multiplying dt and $d\omega_t$ in (5.1) are locally Lipschitz with respect to the supremum norm on sample paths—see Appendix B (that is, given a compact set in the space of ψ_t, $\tilde{\theta}_t$, the functions are Lipschitz with constant depending on the choice of the set).

The result then follows from Theorems (14.18) and (14.20) of [20]. □

Note that Lemma 5.1 does not use the projection of the parameter estimates as described at the end of § 3. When this additional facet of the algorithm is included, we can strengthen Lemma 5.1 as follows.

LEMMA 5.2. *With the addition of the projection scheme* (3.23) *to* (5.1b), *then $\hat{\theta}_t$ remains in $\mathscr{C}_2$ for all t and the composite set of equations* (5.1) *has a unique solution almost surely with sample paths* (ψ_t, $\tilde{\theta}_t$) *which are continuous except at the projection times.*

Proof. Up to the time of the first projection, θ_t is bounded and thus (5.1a) is a linear time varying equation with bounded coefficients and hence ψ_t cannot become unbounded in a finite time. Hence the first projection occurs strictly before the explosion time T of Lemma 5.1 (unless both T and the first projection time are infinite).

After projection, we can apply Lemma 5.1 again and repeat the same argument. Hence due to the linearity of (5.1b) for given $\psi.$, $\tilde{\theta}_t$ exists, is unique and is bounded by the projection. Then (5.1a) is a linear equation with bounded coefficients, and hence ψ_t exists almost surely for all t. □

Lemma 5.2 provides the basic existence and uniqueness result necessary to establish the following result giving properties of the parameter estimator. This result does not depend on a priori boundedness of the system states. Note, however, that we have already established that $\tilde{\theta}_t$ and ψ_t/r_t are bounded.

THEOREM 5.1. *For the general class of feedback control laws described in* § 4 *and under Assumptions* 1 *and* 2, *the following properties hold for the model* ((2.1), (2.2)) *and the estimator* (3.20)-(3.21) *with the projection scheme* (3.23):

$$\text{(i)}\quad \limsup_{t\to\infty}\int_0^t \frac{\eta_\tau^2}{r_\tau}\,d\tau \leqq K_1 < \infty \qquad \textit{a.s.} \tag{5.2}$$

where K_1 is a random variable (*realisation dependent*).

(5.3) (ii) *For all finite* Δ, $\lim_{t\to\infty}\sup_{0\leqq T\leqq\Delta}\|\hat{\theta}_{t+T}-\hat{\theta}_t\| = 0$ *a.s.*

(iii) *There exists a finite random time t_R beyond which no further parameter projections occur.*

Proof. (a) Starting from (3.22) and using Ito's rule (see, e.g., [15]), we have that, between projections:

$$
\begin{aligned}
d(\tilde{\theta}_t^T\tilde{\theta}_t) &= 2\tilde{\theta}_t^T\frac{\psi_t}{r_t}(\eta_t\,dt+d\omega_t)+\sigma^2\frac{\|\psi_t\|^2}{r_t^2}\,dt \\
&= -\frac{2}{r_t}(-\tilde{\theta}_t^T\psi_t\eta_t-\varepsilon\eta_t^2)\,dt-2\varepsilon\frac{\eta_t^2}{r_t}\,dt+2\tilde{\theta}_t^T\frac{\psi_t}{r_t}\,d\omega_t+\sigma^2\frac{\|\psi_t\|^2}{r_t^2}\,dt
\end{aligned}
\tag{5.4}
$$

for some $\varepsilon>0$.

Note that from (3.23) and using Assumption 2, at the times of projection we have $\|\tilde{\theta}_t\|^2 \leqq \|\tilde{\theta}_{t_-}\|^2-(R_2-R_1)^2$.

Defining $\xi_t \triangleq -\tilde{\theta}_t^T\psi_t\eta_t-\varepsilon\eta_t^2$ where ε is as in Assumption 1, integrating (5.4) and accounting for projections yields

$$
\begin{aligned}
\tilde{\theta}_t^T\tilde{\theta}_t \leqq \tilde{\theta}_s^T\tilde{\theta}_s &-2\int_s^t\frac{\xi_\lambda}{r_\lambda}\,d\lambda-2\varepsilon\int_s^t\frac{\eta_\lambda^2}{r_\lambda}\,d\lambda+2\int_s^t\tilde{\theta}_\lambda^T\frac{\psi_\lambda}{r_\lambda}\,d\omega_\lambda \\
&+\int_s^t\sigma^2\frac{\|\psi_\lambda\|^2}{r_\lambda^2}\,d\lambda - N_{t,s}(R_2-R_1)^2
\end{aligned}
\tag{5.5}
$$

where $N_{t,s}$ is the number of times that projections occur between times s and t. Consider now the integral

$$\int_0^t\frac{\|\psi_\lambda\|^2}{r_\lambda^2}\,d\lambda.$$

We have, using (3.21),

$$\int_0^t\frac{\|\psi_\lambda\|^2}{r_\lambda^2}\,d\lambda \leqq \int_0^t\frac{dr_\lambda}{r_\lambda^2}=\frac{1}{r_0}-\frac{1}{r_t}\leqq\frac{1}{r_0}=\frac{1}{c_0}. \tag{5.6}$$

(This operation makes makes sense due to Bonnet's and Du Bois-Reymond's formulae [21] which allow the integral to be considered as a Riemann integral.) We now define $\mathscr{X}_t$ as the solution of the Ito integral

$$d\mathscr{X}_t=\frac{2\tilde{\theta}_t^T\psi_t}{r_t}\,d\omega_t,\qquad \mathscr{X}_0=\frac{\sigma^2}{c_0}+\frac{2K\|\gamma_0\|^2}{r_0}+\tilde{\theta}_0^T\tilde{\theta}_0. \tag{5.7}$$

This integral makes sense thanks to (5.6) and Lemma 5.2. Since c_0, r_0, γ_0, and $\tilde{\theta}_0$ are $\mathscr{F}_0$-measurable, $(\mathscr{X}_t, \mathscr{F}_t)$ is a martingale. Moreover it follows from (5.5) and (5.6) that $\mathscr{X}_t$ satisfies

$$
\begin{aligned}
\mathscr{X}_t \geqq \tilde{\theta}_t^T\tilde{\theta}_t &+2\int_0^t\frac{\xi_\lambda}{r_\lambda}\,d\lambda+2\varepsilon\int_0^t\frac{\eta_\lambda^2}{r_\lambda}\,d\lambda+\frac{\sigma^2}{c_0}-\int_0^t\sigma^2\frac{\|\psi_\lambda\|^2}{r_\lambda^2}\,d\lambda+\frac{2K\|\gamma_0\|^2}{r_0} \\
&+N_{t,0}[R_2-R_1]^2
\end{aligned}
\tag{5.8}
$$

and is positive in view of Lemma 3.1 and (5.6). Thus $(\mathscr{X}_t, \mathscr{F}_t)$ is a positive martingale and hence

$$\lim_{t\to\infty}\mathscr{X}_t=\mathscr{X}<\infty \qquad \text{a.s.} \tag{5.9}$$

Using (5.9) and noting (5.6) and Lemma 3.1, we conclude that (5.2) holds for some finite random variable K_1.

This establishes (i). Also, since $R_2 > R_1$, then from (5.8), (5.9) $N_{t,0}$ is bounded almost surely. Hence (iii) follows.

(b) Now let $t \geqq t_R$. Then from (3.22) we can write, using $\|x+y\|^2 \leqq 2\|x\|^2 + 2\|y\|^2$,

$$\begin{aligned}\|\tilde{\theta}_{t+T}-\tilde{\theta}_t\|^2 &\leqq 2\left\{\left\|\int_t^{t+T}\frac{\psi_\tau\eta_\tau}{r_\tau}\,d\tau\right\|^2+\left\|\int_t^{t+T}\frac{\psi_\tau}{r_\tau}\,d\omega_\tau\right\|^2\right\}\\ &\leqq 2T\int_t^{t+T}\frac{\|\psi_\tau\|^2}{r_\tau^2}\eta_\tau^2\,d\tau+2\left\|\int_t^{t+T}\frac{\psi_\tau}{r_\tau}\,d\omega_\tau\right\|^2.\end{aligned} \tag{5.10}$$

The last inequality is obtained by applying the Schwartz inequality. We consider the two terms of the right-hand side of (5.10) separately.

Consider a realization for which (i) holds. Thus, given $\varepsilon > 0$, $\Delta > 0$ there exists a $t_0(\Delta, \varepsilon)$ such that for all $t \geqq t_0$

$$\int_t^{t+T}\frac{\eta_\tau^2}{r_\tau}\,d\tau \leqq \frac{1}{4}\frac{\varepsilon^2}{\Delta} \quad \forall t \geqq t_0 \quad \text{and} \quad 0 \leqq T \leqq \Delta. \tag{5.11}$$

Since $(\|\psi_\tau\|^2/r_\tau) \leqq 1$ by our definition (3.21) of r_τ, it follows that

$$2T\int_t^{t+T}\frac{\|\psi_\tau\|^2}{r_\tau^2}\eta_\tau^2\,d\tau \leqq \frac{1}{2}\varepsilon^2 \quad \forall t \geqq t_0 \quad \text{and} \quad \forall\, T \in [0, \Delta] \tag{5.12}$$

and hence

$$\limsup_{t\to\infty}\ \sup_{0\leqq T\leqq\Delta} 2T\int_t^{t+T}\frac{\|\psi_\tau\|^2}{r_\tau^2}\eta_\tau^2\,d\tau = 0. \tag{5.13}$$

Since (i) holds almost surely, so does (5.13). The result then follows using part (a) of Lemma C.1 (Appendix C) for the second term on the right-hand side of (5.10). □

To make use of the result in Theorem 5.1, we will need the following continuous-time Kronecker lemma. Since we have been unable to find a proof in the literature for this continuous-time version, we supply one in Appendix D. It parallels the discrete-time proof given in [2].

LEMMA 5.3 (Continuous-time Kronecker Lemma). *Assume that*

(A.1) $S_t = \int_0^t x_\tau\,d\tau$ *converges to* $S < \infty$ *as* $t \to \infty$.

(A.2) $b_t > 0$ *is monotone nondecreasing and* $\lim_{t\to\infty} b_t = \infty$.

Then

$$\lim_{t\to\infty}\frac{1}{b_t}\int_0^t b_\tau x_\tau\,d\tau = 0.$$

Proof. For the proof see Appendix D.

We now prove a continuous time equivalent of the discrete-time key technical lemma given in [1], [2].

LEMMA 5.4. *Consider a realisation produced by the model* (2.1)–(2.2). *Suppose the estimator is such that*

$$\limsup_{t\to\infty}\int_0^t\frac{\eta_\tau^2}{r_\tau}\,d\tau \leqq K_1 < \infty \tag{5.14}$$

and the controller is such that the following growth condition is satisfied:

$$\frac{r_t}{t} \leqq C + \frac{K_2}{t}\int_0^t \eta_\tau^2\,d\tau \tag{5.15}$$

where r_t, η_t are as defined before and C, K_2 are finite positive constants. Then

$$(5.16) \qquad \text{(i)} \ \limsup_{t\to\infty} \frac{r_t}{t} \leqq K_3 < \infty$$

$$(5.17) \qquad \text{(ii)} \ \lim_{t\to\infty} \frac{1}{t} \int_0^t \eta_\tau^2 \, d\tau = 0.$$

Proof. (i) From (5.15) and the nonnegativity of η_τ^2 and r_τ it follows that

$$\frac{r_t}{t} \leqq C + K_2 \int_0^t \frac{\eta_\tau^2}{r_\tau} \frac{r_\tau}{\tau} \, d\tau.$$

The result then follows from the Bellman-Gronwall lemma (see e.g., [17]) using (5.14):

(ii) Suppose first that $\lim_{t\to\infty} r_t = \infty$. Then, by the Kronecker Lemma 5.3:

$$(5.18) \qquad \lim_{t\to\infty} \frac{1}{r_t} \int_0^t \eta_\tau^2 \, d\tau = 0.$$

Hence, using (5.16) and (5.18)

$$(5.19) \qquad \lim_{t\to\infty} \frac{1}{t} \int_0^t \eta_\tau^2 \, d\tau = \lim_{t\to\infty} \frac{r_t}{t} \frac{1}{r_t} \int_0^t \eta_\tau^2 \, d\tau = 0.$$

Alternatively, if $\lim_{t\to\infty} r_t \leqq K_4 < \infty$ for some K_4, then

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \eta_\tau^2 \, d\tau \leqq \lim_{t\to\infty} \frac{1}{t} \frac{K_4}{r_t} \int_0^t \eta_\tau^2 \, d\tau$$

$$\leqq \lim_{t\to\infty} \frac{K_4}{t} \int_0^t \frac{\eta_\tau^2}{r_\tau} \, d\tau = 0$$

using (5.2). The last inequality follows from the monotonicity of r_τ. □

Comment 5.1. A key feature of our estimator is our choice of r_t (see (3.21)). It guarantees that $(\|\psi_t\|^2/r_t) \leqq 1$ for all t, and this is crucial in establishing (5.2). In [5], [6] the equivalent term to r_t is defined as $r_t = \int_0^t \psi_\tau^T \psi_\tau \, d\tau$ and, as a consequence, all the convergence proofs require the assumption that $(\|\psi_t\|^2/r_t)$ is almost surely bounded. This assumption is rather unrealistic given that ψ_t contains signals driven by white noise.

Comment 5.2. Comparing the properties of our estimator with those proved for the continuous-time deterministic algorithms in [9], we observe that we have not proved the uniform boundedness of $(\eta_t/r_t^{1/2})$ or of something like $\dot{\hat{\theta}}$; neither have we proved $\dot{\hat{\theta}} \in L_2$. The uniform boundedness of $(\eta_t/r_t^{1/2})$ is not needed in the subsequent analysis $((\eta_t/r_t^{1/2}) \in L_2$ suffices). As for $\dot{\tilde{\theta}}$, it does not exist in our stochastic framework, but the conditions on $\dot{\tilde{\theta}}$ in [9] are replaced by the weaker condition of (5.3).

Comment 5.3. The remaining step in the development is to verify the growth condition (5.15). This will require us to impose additional constraints on the general feedback law described in § 4. In particular, we will choose the feedback so as to stabilize the (frozen) estimated model.

6. Boundedness of the system states. Finally we show that, provided the feedback control law is appropriately chosen as a function of $\hat{\theta}$, then the adaptive law stabilizes the system in the sense that all system states and the input are mean square bounded almost surely.

The proofs given below depend upon the fact that the certainty equivalence control law stabilize the frozen *estimated* model. This is true of a wide class of algorithms. However, to be specific, we will illustrate the analysis procedure by considering adaptive pole assignment. In this algorithm the polynomials $\hat{L}$ and $\hat{P}$ defining the control law are computed from the estimated $\hat{A}(\hat{\theta}_t)$ and $\hat{B}(\hat{\theta}_t)$ as follows.

Let

$$\hat{\theta} = [\hat{g}_{n-1}, \cdots, \hat{g}_0, \hat{b}_{n-1}, \cdots, \hat{b}_0, \hat{f}_{n-1}, \cdots, \hat{f}_0]$$

and define (cf. (2.6), (2.21))

$$\begin{aligned} &\hat{a}_i = e_i - \hat{g}_i; \qquad i = 0, \cdots, n-1 \\ &\hat{A}(\rho) = \rho^n + \hat{a}_{n-1}\rho^{n-1} + \cdots + \hat{a}_0 \\ &\hat{B}(\rho) = \hat{b}_{n-1}\rho^{n-1} + \cdots + \hat{b}_0 \\ &\hat{F}(\rho) = \hat{f}_{n-1}\rho^{n-1} + \cdots + \hat{f}_0. \end{aligned}$$

Then, for a given possibly time varying A^* of degree $2n$, solve the following equation for $\hat{Q}$ and $\hat{P}$:

$$\hat{Q}\hat{A} + \hat{P}\hat{B} = A^* \quad \text{with} \quad \operatorname{Re} \lambda_i(A^*) \leqq -\beta < 0, \qquad i = 1, \cdots, 2n \tag{6.1}$$

where $\lambda_i(A^*)$ are the eigenvalues of the polynomial A^*. Finally compute

$$\hat{L}(\rho) = \hat{Q}(\rho) - E(\rho).$$

Equation (6.1) can also be written

$$M(\hat{\theta}) \begin{bmatrix} \hat{q}_0 \\ \vdots \\ \hat{q}_{n-1} \\ \hat{p}_0 \\ \vdots \\ \hat{p}_{n-1} \end{bmatrix} = \begin{bmatrix} a_0^* \\ \vdots \\ \vdots \\ \vdots \\ a_{2n-1}^* \end{bmatrix}. \tag{6.2}$$

The polynomial A^* is a design polynomial available to the user. The only restriction we require on the polynomial $A^*(t)$ is seen in the following assumption.

Assumption 3. For all t, the polynomial $A^*(t)$ is continuous and bounded, has a uniform stability margin (i.e., $\operatorname{Re} \lambda_i(A^*(t)) \leqq -\beta < 0$, $i = 1, \cdots, 2n$), and $\lim_{t\to\infty} \sup_{0 \leqq T \leqq \Delta} \|a^*(t+T) - a^*(t)\| = 0$ for some $\Delta > 0$, where a^* is the vector of coefficients of A^*.

As in § 4, we also require that $\hat{L}$, $\hat{P}$ be Lipschitz functions of $\hat{\theta}$ for all time. We note that the projection facility (3.23) ensures that $\hat{\theta} \in \mathscr{C}_2$ for all time. Then, the Lipschitz condition is guaranteed provided all models corresponding to $\hat{\theta} \in \mathscr{C}_2$ are uniformly stabilizably. This assumption appears in all contemporary treatments of indirect adaptive control (see, for example, [2], [10], [19]). We remark that this assumption can be eliminated in the case of direct adaptive control of stably invertible systems. However, this is only achieved at the expense of a much more complex stability proof (see [9], [22]) which necessarily builds on the result presented here. For the case of indirect adaptive pole assignment we introduce the following additional assumption.

Assumption 4. Assumption 2 is satisfied and there exists a known positive constant ε such that for all θ_t in $\mathscr{C}_2$,

$$\det M(\theta_t) \geqq \varepsilon.$$

Subject to Assumption 4, the projection scheme ensures that $\det M(\hat{\theta}_t) \geqq \varepsilon$ for all t. In view of (6.2) this will ensure that $\hat{L}$, $\hat{P}$ are Lipschitz functions of $\hat{\theta}$ as required.

We next establish that, for the above adaptive control law, the homogeneous part of (4.5) is exponentially stable.

LEMMA 6.1. *Consider the differential equation*

$$\frac{d}{dt}\psi_t = A_t\psi_t$$

with A_t given by (4.6a). *Assume that the $\hat{a}_i$, $\hat{b}_i$, $\hat{f}_i$ are estimated using the parameter estimator of* § 3 *including the projection scheme* (3.23) *and that Assumptions* 1–4 *hold. Then* (6.1) *is exponentially stable almost surely.*

Proof. We have shown in Theorem 5.1 that there exists a random time t_R beyond which no further projections occur. Thus, for $t > t_R$ $\hat{\theta}_t$ is sample continuous from Lemma 5.2. Therefore $\hat{A}$, $\hat{B}$, $\hat{F}$ are sample continuous and, by Assumption 4, $\hat{Q}$ and $\hat{P}$ are sample continuous given that $\hat{q}_i$ and $\hat{p}_i$ can be written as the solution of the linear system (6.2) or $M(\hat{\theta})\hat{c} = a^*$, where the elements of M are either zero or $\hat{a}_i$, $\hat{b}_j$. For the same reason, A_t is uniformly bounded. The eigenvalues of A_t are the roots of $E(s)$ and the $2n$ roots of $\hat{A}^*$; therefore, by Assumptions 1 and 3 $\operatorname{Re} \lambda_i(A_t) \leqq -\delta < 0$, $i = 1, \cdots, 3n$, for all t, where $\delta \triangleq \min(\alpha, \beta)$. Finally,

$$M(\hat{\theta}_t)[\hat{c}_{t+T} - \hat{c}_t] + [M(\hat{\theta}_{t+T}) - M(\hat{\theta}_t)]\hat{c}_{t+T} = a^*_{t+T} - a^*_t.$$

Therefore, by the triangle inequality,

$$\|\hat{c}_{t+T} - \hat{c}_t\| \leqq \|M^{-1}(\hat{\theta}_t)\| \{\|M(\hat{\theta}_{t+T}) - M(\hat{\theta}_t)\| \cdot \|\hat{c}_{t+T}\| + \|a^*_{t+T} - a^*_t\|\}.$$

Hence, using (5.3) and Assumption 3 it follows that

$$\lim_{t\to\infty} \sup_{0 \leqq T \leqq \Delta} \|A_{t+T} - A_t\| = 0 \quad \text{a.s. for some } \Delta > 0. \tag{6.3}$$

The result then follows from Lemma 3 of [10]. □

We now establish the main result of this paper.

THEOREM 6.1. *Consider the system* ((2.1), (2.2)), *with the parameter vector θ satisfying Assumption* 2, *the parameter estimator of* § 3 *with projection, an observer polynomial E satisfying Assumption* 1, *and an adaptive pole assignment control law of the form* (4.2), (6.1) *satisfying Assumptions* 3 *and* 4. *Then, for arbitrary finite initial conditions and an arbitrary, piecewise continuous, uniformly bounded reference input y^*_t, the following results hold*:

(i) $\displaystyle \limsup_{t\to\infty} \frac{1}{t}\int_0^t \|\psi_\tau\|^2 \, d\tau < \infty$ *a.s.*

(ii) $\displaystyle \limsup_{t\to\infty} \frac{1}{t}\int_0^t u_\tau^2 \, d\tau < \infty$ *a.s.*

(iii) $\displaystyle \limsup_{t\to\infty} \frac{1}{t}\int_0^t |Cx_\tau|^2 \, d\tau < \infty$ *a.s.*

Proof. (i) We first establish (5.15), i.e., that there exist finite random variables C and K such that:

$$\frac{r_t}{t} \leqq C + \frac{K}{t}\int_0^t \eta_\tau^2 \quad \text{a.s.}$$

Now by definition

$$r_t \triangleq \sup_{0 \leqq \tau \leqq t} \|\psi_\tau\|^2 + \int_0^t \|\psi_\tau\|^2 \, d\tau + c_0. \tag{6.4}$$

From (4.5)

$$\psi_t = \phi(t,0)\psi_0 + \int_0^t \phi(t,\tau)B_1\eta_\tau\,d\tau + \int_0^t \phi(t,\tau)B_2 y_\tau^*\,d\tau + \int_0^t \phi(t,\tau)B_1\,d\omega_\tau$$

$$= \psi_t^{(1)} + \psi_t^{(2)} \tag{6.5}$$

where $\psi_t^{(1)}$ is the sum of the first three terms, $\psi_t^{(2)}$ is the last term and $\phi(t,\tau)$ denotes the state transition matrix for (4.5). Because ϕ is exponentially stable by Lemma 6.1 and y_τ^* is uniformly bounded, it follows using the Cauchy-Schwartz inequality that there exist finite constants K_1 and K_2 such that

$$\|\psi_t^{(1)}\|^2 \leqq K_1 + K_2 \int_0^t \eta_\tau^2\,d\tau \tag{6.6}$$

for all t. Also, by Lemma C.2

$$\frac{1}{t}\|\psi_t^{(2)}\|^2 \to 0 \quad \text{a.s.} \tag{6.7}$$

as $t \to \infty$.

Similarly, using a continuous-time version of the proof of Lemma B.3.3 in [2], it follows that there exist finite constants K_3 and K_4 such that

$$\int_0^t \|\psi_\tau^{(1)}\|^2\,d\tau \leqq K_3 + K_4 \int_0^t \eta_\tau^2\,d\tau \tag{6.8}$$

for all t. Finally, by Lemma C.2,

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t \|\psi_\tau^{(2)}\|^2\,d\tau \leqq K_5 < \infty \quad \text{a.s.} \tag{6.9}$$

Combining (6.6) to (6.9) we obtain

$$\begin{aligned}\frac{r_t}{t} &= \frac{1}{t}\sup_{0\leqq\tau\leqq t}\|\psi_\tau\|^2 + \frac{1}{t}\int_0^t \|\psi_\tau\|^2\,d\tau + \frac{c_0}{t}\\ &\leqq \frac{1}{t}\max\left[\sup_{0\leqq\tau\leqq 1}\|\psi_\tau\|^2,\ \sup_{1\leqq\tau\leqq t}\|\psi_\tau\|^2\right] + \frac{1}{t}\int_0^t \|\psi_\tau\|^2\,d\tau + \frac{c_0}{t}\\ &\leqq K_6 + \frac{K_7}{t}\int_0^t \eta_\tau^2\,d\tau \quad \text{a.s.}\end{aligned} \tag{6.10}$$

Thus, we have established (5.15) so that, by Lemma 5.4, there exists a finite K_8 such that

$$\frac{r_t}{t} \leqq K_8 \quad \text{a.s.} \tag{6.11}$$

It follows from the definition of r_t that

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t \|\psi_\tau\|^2\,dt < \infty \quad \text{a.s.,} \tag{6.12}$$

which establishes part (i) of the theorem.

(ii) Since $u_t = \hat{\xi}_t\psi_t + y_t^*$, where $\hat{\xi}_t$ is a vector whose components are $\{\hat{p}_i\}$ and $\{\hat{l}_i\}$ (see (4.4) and the definition of ψ_t) it follows from (6.10) and the boundedness of $\{\hat{p}_i\}$ and $\{\hat{l}_i\}$ that

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t u_\tau^2\,d\tau < \infty \quad \text{a.s.} \tag{6.13}$$

which establishes part (ii) of the theorem.

(iii) From (2.9), (2.19), (3.7), and (3.8) we have

$$Cx_t = \psi_t^T \hat{\theta}_t + \eta_t. \tag{6.14}$$

Therefore, since $\hat{\theta}_t$ is bounded, $\lim_{t\to\infty} \sup (1/t) \int_0^t \eta_t^2 \, d\tau = 0$, (Lemma 5.4) and $\lim_{t\to\infty} \sup (1/t) \int_0^t \|\psi_\tau\|^2 \, d\tau < \infty$ almost surely (see (6.12)), it follows that

$$\limsup_{t\to\infty} \frac{1}{t} \int_0^t [Cx_t]^2 \, dt < \infty \quad \text{a.s.} \tag{6.15}$$

which establishes part (iii) of the theorem. □

Comment 6.1. Note that Cx_t is the "deterministic part" of the output. Thus, (6.15) establishes that this part of the output is sample mean square bounded. This is all that can be said about the output since it contains a Wiener process.

Comment 6.2. The proof of Theorem 6.1 relies on the now standard argument on robustness of bounded solutions of exponentially stable linear time varying systems.

7. Conclusions. This paper has analyzed a class of continuous-time stochastic adaptive control algorithms and has shown that, under suitable conditions, they will almost surely ensure global boundedness of all the internal variables, in a sample mean square sense. Previous results in the same direction relied upon an assumption on the normalized regression vector which, in fact, meant that the noise driving the system was wide band but bounded. Moreover, the systems described in previous works were of relative degree zero.

The results described in this paper are thus believed to constitute the first complete and rigorous analysis of a realistic continuous time stochastic adaptive control algorithm.

The main result is a boundedness result that states that the deterministic part of the output remains almost surely bounded in the sample mean square sense. No tracking property is achieved, but the analysis described here has recently been extended to a model reference control algorithm yielding a result on the asymptotic tracking error [22], which is analogous to the discrete-time tracking error result established in [1].

Appendix A.

LEMMA A.1. *Let r_τ be a nondecreasing function with $r_0 > 0$. Then*

$$\int_0^T f_\tau \, d\tau \geqq -K; \qquad K > 0, \quad \forall T \geqq 0 \tag{A.1}$$

implies

$$\int_0^T \frac{f_\tau}{r_\tau} \, d\tau \geqq -\frac{K}{r_0}; \quad \forall T \geqq 0. \tag{A.2}$$

Proof. Let $\rho_\tau = (1/r_\tau)$, $\tau \geqq 0$, and denote $F_t = \int_0^t f_\tau \, d\tau + K$. Then, by integration by parts:

$$\int_0^T f_\tau \rho_\tau \, d\tau = \rho_T F_T + \left\{ -\int_0^T F_\tau \, d\rho_\tau \right\} - \rho_0 K.$$

The first term is nonnegative because $\rho_T > 0$ and $F_T \geqq 0$. The second term is also nonnegative because $F_\tau \geqq 0$, for all $\tau \geqq 0$ by (A.1), and $d\rho_\tau \leqq 0$ since ρ_τ is nonincreasing. □

Appendix B.

LEMMA B.1. *The coefficient vectors in* (5.1) *are locally Lipschitz with respect to the supremum norm on the sample paths ψ_t, $\tilde{\theta}_t$.*

Proof. We consider (5.1a), (5.1b) separately. Equation (5.1a) has the form:

$$d\psi_t = f_1(t, \tilde{\theta}_\cdot, \psi_\cdot)\,dt + B_1\,d\omega_t + B_2 y_t^*\,dt \tag{B.1}$$

where

$$f_1(t, \tilde{\theta}_\cdot, \psi_\cdot) = \left[A_t(\tilde{\theta}_t)\psi_t - B_1\psi_t^T\tilde{\theta}_t - B_1\int_0^t h_{t-\tau}\psi_\tau^T\tilde{\theta}_\tau\,d\tau\right]. \tag{B.2}$$

We then have, by the triangle inequality,

$$\begin{aligned}
&\|f_1(t, \tilde{\theta}_\cdot^1, \psi_\cdot^1) - f_1(t, \tilde{\theta}_\cdot^2, \psi_\cdot^2)\| \\
&\leqq \|A_t(\tilde{\theta}_t^1)\psi_t^1 - A_t(\tilde{\theta}_t^2)\psi_t^2\| + \|-B_1(\psi_t^1)^T\tilde{\theta}_t^1 + B_1(\psi_t^2)^T\tilde{\theta}_t^2\| \\
&\quad + \left\|-B_1\int_0^t h_{t-\tau}\left[(\psi_\tau^1)^T\tilde{\theta}_\tau^1 - (\tilde{\psi}_\tau^2)^T\tilde{\theta}_\tau^2\right]d\tau\right\| \\
&\leqq \|A_t(\tilde{\theta}_t^1) - A_t(\tilde{\theta}_t^2)\|\ \mathrm{Max}\,(\|\psi_t^1\|, \|\psi_t^2\|) \\
&\quad + \mathrm{Max}\,(\|A_t(\tilde{\theta}_t^1)\|, \|A_t(\tilde{\theta}_t^2)\|)\|\psi_t^1 - \psi_t^2\| + \|B_1\|\,\|\psi_t^1 - \psi_t^2\|\ \mathrm{Max}\,(\|\tilde{\theta}_t^1\|, \|\tilde{\theta}_t^2\|) \\
&\quad + \|B_1\|\ \mathrm{Max}\,(\|\psi_t^1\|, \|\psi_t^2\|)\|\tilde{\theta}_t^1 - \tilde{\theta}_t^2\| + \|B_1\|\left(\int_0^t |h_{t-\tau}|\,d\tau\right)
\end{aligned} \tag{B.3}$$

$$\begin{aligned}
&\left\{\sup_{0\leqq\tau\leqq t}\|\psi_\tau^1 - \psi_\tau^2\|\ \sup_{0\leqq\tau\leqq t}\mathrm{Max}\,(\|\tilde{\theta}_\tau^1\|\cdot\tilde{\theta}_\tau^2\|)\right. \\
&\left.\quad + \sup_{0\leqq\tau\leqq t}\mathrm{Max}\,(\|\psi_\tau^1\|, \|\psi_\tau^2\|)\ \sup_{0\leqq\tau\leqq t}\|\tilde{\theta}_\tau^1 - \tilde{\theta}_\tau^2\right\} \\
&\leqq K_n\left\{\sup_{0\leqq\tau\leqq t}\|\psi_\tau^1 - \psi_\tau^2\| + \sup_{0\leqq\tau\leqq t}\|\tilde{\theta}_\tau^1 - \tilde{\theta}_\tau^2\|\right\}
\end{aligned} \tag{B.4}$$

where the last line follows since $A_t(\tilde{\theta}_t)$ is Lipschitz in $\tilde{\theta}_t$ and since $\int |h_{t-\tau}|\,d\tau < \infty$ by the exponential stability of $(E-D)/D$. In (B.4), the notation K_n indicates that the Lipschitz constant depends on the maximum values of ψ, $\tilde{\theta}$ in the compact set defining the local conditions.

For (5.1b) the proof is similar on noting that, by the Cauchy-Schwartz inequality,

$$\frac{|\psi_t^T\psi_{t-\tau}|}{r_t} \leqq \frac{|\psi_t^T\psi_t|^{1/2}|\psi_{t-\tau}^T\psi_{t-\tau}|^{1/2}}{r_t} \tag{B.5}$$

and that

$$\frac{|\psi_t^T\psi_t|}{r_t} \leqq 1 \quad\text{and}\quad \frac{|\psi_{t-\tau}^T\psi_{t-\tau}|}{r_t} \leqq 1 \quad \text{by definition of } r_t. \qquad \square$$

Appendix C.

LEMMA C.1 ("In-flight Lemma"). *Let* $S_t = \int_0^t h(\tau)\,d\omega_\tau$. *Assume that*

(C.1)
(i) $h(\tau)$ *is* $\mathcal{F}_s$ *measurable for* $\tau \leqq s$, *and*
(ii) $\int_0^\infty \|h(\tau)\|^2\,d\tau \leqq K < \infty$ a.s.
where K *is* $\mathcal{F}_0$*-measurable.*

Then

(C.2)
(a) *given* $\Delta > 0$,
$\lim_{t\to\infty}\sup_{0\leqq\tau\leqq\Delta}\|S_{t+\tau} - S_t\|^2 = 0$ a.s.
(b) $\|S_t\|^2$ *converges to a finite limit* a.s.

Proof. (a) Take any $t_0 \in [0, \Delta)$ and partition the positive real line in intervals of length Δ: $0, t_0, t_0+\Delta, t_0+2\Delta, \cdots$. Define

$$L_j(t_0) = \sup_{T \in [0,\Delta]} \|S_{t_0+(j-1)\Delta+T} - S_{t_0+(j-1)\Delta}\|^2; \qquad j = 1, 2, \cdots \tag{C.3}$$

and let

$$T_j(t_0) = \arg\left\{ \sup_{T \in [0,\Delta]} \|S_{t_0+(j+1)\Delta+T} - S_{t_0+(j-1)\Delta}\|^2 \right\}. \tag{C.4}$$

Also define

$$J_n(t_0) = \sum_1^n L_j + K - \sum_{j=1}^n \int_{t_0+(j-1)\Delta}^{t_0+(j-1)\Delta+T_j} \|h(\tau)\|^2 \, d\tau; \qquad n = 1, 2, \cdots. \tag{C.5}$$

Then, by assumption (i), we have

$$\begin{aligned} E\{J_{n+1}(t_0) \mid \mathscr{F}_n\} &= J_n(t_0) + E\{L_{n+1}(t_0) \mid \mathscr{F}_n\} - E\left\{ \int_{t_0+n\Delta}^{t_0+n\Delta+T_n} \|h(\tau)\|^2 \, d\tau \mid \mathscr{F}_n \right\} \\ &= J_n(t_0) \quad \text{a.s.} \end{aligned}$$

Therefore, using assumption (ii), $J_n(t_0)$ is a nonnegative martingale, and hence converges almost surely to a finite limit [16]. Hence, since the sum of the last two terms in (C.5) decreases monotonically, $L_j(t_0)$ converges to zero almost surely. Now recalling (C.3), since $L_j(t_0)$ goes to zero (almost surely) for any t_0, we conclude that (a) holds.

(b) By the Ito rule [13]

$$\|S_t\|^2 = \|S_s\|^2 + 2S_s \int_s^t h(\tau) \, d\omega_\tau + \sigma^2 \int_s^t |h(\tau)\|^2 \, d\tau. \tag{C.6}$$

Now define

$$X_t = \|S_t\|^2 + K\sigma^2 - \sigma^2 \int_0^t \|h(\tau)\|^2 \, d\tau. \tag{C.7}$$

We note that X_t is positive and $\mathscr{F}_s$-measurable for $t \leqq s$. Substituting (C.6) into (C.7)

$$\begin{aligned} X_t &= \|S_s\|^2 + 2S_s \int_s^t h(\tau) \, d\omega_\tau + \sigma^2 \int_s^t \|h(\tau)\|^2 \, d\tau + K\sigma^2 - \sigma^2 \int_0^t \|h(\tau)\|^2 \, d\tau \\ &= X_s + 2S_s \int_s^t h(\tau) \, d\omega_\tau. \end{aligned} \tag{C.8}$$

Taking conditional expectations yields

$$E[X_t \mid \mathscr{F}_s] = X_s. \tag{C.9}$$

Therefore, using (C.1) in (C.7), X_t is a nonnegative martingale and hence converges [16] to a finite limit almost surely. This limit is a random variable. The last term in (C.7) is monotone nondecreasing and bounded, thus it converges also. Thus $\|S_t\|^2$ converges also to a finite random variable. □

LEMMA C.2. *Let $S_t \triangleq \int_0^t \phi(t, \tau)a_\tau \, d\omega_\tau$ where $\phi(t, \tau)$ is the state transition matrix of an exponentially stable system and a_τ is uniformly bounded. Then, there exist finite random variables K_1 and K_2, such that*

(C.10) (i) $\limsup_{t\to\infty} \|S_t\|^2 < K_1$ a.s.

(C.11) (ii) $\limsup_{t\to\infty} \frac{1}{t}\int_0^t \|S_t\|^2 \, d\tau < K_2$ a.s.

Proof. (i) $S_t = S_s + \int_s^t \phi(t, \tau)a_\tau \, d\omega_\tau$. Hence using the Ito rule we have

$$S_t^2 = S_s^2 + 2S_s \int_s^t \phi(t, \tau)a_\tau \, d\omega_\tau + \int_s^t \sigma^2\phi(t, \tau)^2 a_\tau^2 \, d\tau.$$

Let $\mathcal{F}_t$ denote the increasing σ-fields generated by $\{\omega_s, 0 \leqq s \leqq t\}$. Then

$$E\{S_t^2 | \mathcal{F}_s\} = S_s^2 + \sigma^2 \int_s^t \phi(t, \tau)^2 a_\tau^2 \, d\tau.$$

We note that, by the assumption on ϕ and a, there exists a constant $K < \infty$ such that

$$\int_0^T \phi(T, \tau)^2 a_\tau^2 \, d\tau < K \text{ for all } T.$$

Let X_t be defined by

$$X_t \triangleq S_t^2 + \sigma^2 K - \sigma^2 \int_0^t \phi(t, \tau)^2 a_\tau^2 \, d\tau.$$

Clearly $X_t \geqq 0$ for all t and

$$E[X_t | \mathcal{F}_s] = S_s^2 + \sigma^2 K - \sigma^2 \int_0^s \phi(s, \tau)^2 a_\tau^2 \, d\tau$$
$$= X_s.$$

It follows that $(X_t, \mathcal{F}_t)$ is a positive martingale so that there exists a finite random variable X_∞ such that

$$X_t \to X_\infty \quad \text{a.s.}$$

as $t \to \infty$. Hence

$$\limsup_{t\to\infty} S_t^2 \leqq X_\infty \quad \text{a.s.}$$

(ii) This part follows as in part (i). □

Appendix D.

CONTINUOUS TIME KRONECKER LEMMA. *Assume that*

(D.1) $S_t = \int_0^t x_\tau \, d\tau$ *converges to* $S < \infty$ *as* $t \to \infty$

(D.2) $b_t \geqq 0$ *is monotone nondecreasing and* $\lim_{t\to\infty} b_t = \infty$.

Then

$$\lim_{t\to\infty} \frac{1}{b_t} \int_0^t b_\tau x_\tau \, d\tau = 0.$$

Proof. (1) We first establish that, under the same assumptions:

$$\lim_{t\to\infty} y_t = 0, \text{ where } y_t \triangleq \frac{1}{b_t}\int_0^t (S_\tau - S)\, db_\tau.$$

By (D.1), for any $\varepsilon > 0$, $\exists t(\varepsilon)$ s.t. $\forall \tau \geqq t(\varepsilon)$, $|S_\tau - S| < \varepsilon$. Therefore

$$|y_t| \leqq \frac{1}{b_t}\left|\int_0^{t(\varepsilon)} (S_\tau - S)\, db_\tau\right| + \frac{1}{b_t}\int_{t(\varepsilon)}^t |S_\tau - S|\, db_\tau$$
$$\leqq \frac{1}{b_t}\cdot C(t(\varepsilon)) + \varepsilon\left[1 - \frac{b(t(\varepsilon))}{b_t}\right]$$

where

$$\frac{b(t(\varepsilon))}{b_t} < 1, \lim_{t\to\infty}\frac{b(t(\varepsilon))}{b_t} = 0$$

and $C(t(\varepsilon))$ is a constant. Therefore by (D.2), $\lim_{t\to\infty}|y_t| \leqq \varepsilon$ with ε arbitrarily small, and hence $\lim_{t\to\infty} y_t = 0$.

(2) Integrating by parts,

$$\frac{1}{b_t}\int_0^t b_\tau x_\tau\, d\tau = \frac{1}{b_t}[b_\tau S_\tau]_0^t - \frac{1}{b_t}\int_0^t S_\tau\, db_\tau$$
$$= S_t - \frac{1}{b_t}\int_0^t S_\tau\, db_\tau \quad (\text{using } S_0 = 0)$$
$$= S_t - S - \frac{1}{b_t}\int_0^t (S_\tau - S)\, db_\tau.$$

Using (D.1) and part (1) of the proof establishes the result. □

Acknowledgments. The authors gratefully acknowledge technical help during the writing of this paper from Bob Bitmead, David Mayne, Sean Meyn, and Peter Moylan. In addition, the authors are particularly grateful to Rene Boel and Mark Davies for their assistance with the proof of existence and uniqueness of the equations describing the system and algorithm.

REFERENCES

[1] G. C. Goodwin, P. J. Ramadge, and P. E. Caines, *Discrete time stochastic adaptive control,* SIAM J. Control Optim., 19 (1981), pp. 829-853.

[2] G. C. Goodwin and K. S. Sin, *Adaptive Filtering Prediction and Control,* Prentice Hall, Englewood Cliffs, NJ, 1984.

[3] P. R. Kumar, *A survey of some results in stochastic adaptive control,* SIAM J. Control Optim., 23 (1985), pp. 329-380.

[4] J. H. Van Schuppen, *Convergence results for continuous time adaptive stochastic filtering algorithms,* J. Math. Anal. Appl., 96 (1983), pp. 209-225.

[5] H. F. Chen and L. Guo, *Continuous-time stochastic adaptive tracking: robustness and asymptotic properties,* Technical Report 1987, Institute of Systems Science, Academia Sinica, Beijing, China.

[6] H. F. Chen, *Recursive Estimation and Control for Stochastic Systems,* John Wiley 1985, New York.

[7] J. B. Moore, *Convergence of continuous time stochastic* ELS *parameter estimation,* Internal Report, Department of Systems Engineering, Australian National University, Australia, 1986.

[8] H. F. Chen and J. B. Moore, *Convergence rates of continuous time stochastic* ELS *parameter estimation,* IEEE Trans. Automat. Control, AC-31 (1987), pp. 267-269.

[9] G. C. Goodwin and D. Q. Mayne, *A parameter estimation perspective of continuous time model reference adaptive control*, Automatica, 23 (1987), pp. 57-70.

[10] G. Kreisselmeier, *An approach to stable indirect adaptive control*, Automatica, 21 (1985), pp. 425-433.

[11] R. H. Middleton, G. C. Goodwin, D. J. Hill, and D. Q. Mayne, *Design issues in adaptive control*, IEEE Trans. Automat. Control, 33 (1988), pp. 50-58.

[12] K. J. Astrom, *Introduction to Stochastic Control Theory*, Academic Press, New York, 1970.

[13] V. Solo, *Topics in Advanced Time Series Analysis*, Springer Lecture Notes, Springer-Verlag, New York, to appear.

[14] I. I. Gihman and A. V. Skorohod, *Controlled Stochastic Processes*, Springer-Verlag, New York, 1979.

[15] E. Wong, *Stochastic Processes in Information and Dynamical Systems*, McGraw Hill, New York, 1971.

[16] J. L. Doob, *Stochastic Processes*, John Wiley, New York, 1953.

[17] C. A. Desoer and M. Vidyasagar, *Feedback Systems: Input-Ouput Properties*, Academic Press, New York, 1975.

[18] V. Wertz, G. C. Goodwin, H. F. Chen, and M. Gevers, *Unification of discrete and continuous time stochastic adaptive control algorithms*, 8th IFAC/IFORS Symp. on Identification and System Parameter Estimation, Vol. 1, Beijing, China, Aug. 1988, pp. 121-126.

[19] Ph. de Larminat, *Explicit adaptive control without persistently exciting inputs*, 2nd IFAC Workshop on Adaptive Systems in Control and Signal Processing, Lund, Sweden, 1986.

[20] J. Jacod, *Calcul Stochastique et Problemes de Martingales*, Lecture Notes in Mathematics 714, Springer-Verlag, Berlin, New York, 1979.

[21] P. Deheuvels, *L'integrale*, Presses Universitaires de France, Mathematiques, 1980.

[22] G. C. Goodwin and D. Q. Mayne, *Continuous time stochastic model reference adaptive control*, IEEE Trans. Automat. Control, 1990, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 283-324, March 1991

003

A CHARACTERIZATION OF ALL SOLUTIONS TO THE FOUR BLOCK GENERAL DISTANCE PROBLEM*

K. GLOVER†, D. J. N. LIMEBEER‡, J. C. DOYLE§, E. M. KASENALLY‡, AND M. G. SAFONOV¶

Abstract. All solutions to the four block general distance problem which arises in H^∞ optimal control are characterized. The procedure is to embed the original problem in an all-pass matrix which is constructed. It is then shown that part of this all-pass matrix acts as a generator of all solutions. Special attention is given to the characterization of all optimal solutions by invoking a new descriptor characterization of all-pass transfer functions. As an application, necessary and sufficient conditions are found for the existence of an H^∞ optimal controller. Following that, a descriptor representation of all solutions is derived.

Key words. H^∞-optimal control, four block problem, Parrott's theorem, general distance problems, indefinite Riccati equations, indefinite factorization, linear quadratic differential games, Nehari's theorem

AMS(MOS) subject classifications.

1. Introduction. The four block general distance problem has its genesis in certain recent work on H^∞ optimal control [7], [10], [14], [15], [43]. In typical H^∞ design situations, a nominal plant model is known, and the design engineer has the task of selecting various frequency-dependent weights. The plant model and weights are then combined into a single matrix

$$P(s)=\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}(s) \tag{1.1}$$

and we seek to characterize all internally stabilizing controllers which satisfy $\|F_l(P, K)\|_\infty \leqq \gamma$. Rather than tackling this (nonlinear) problem directly, it can be converted into another problem which is linear in a *free parameter*. That is, find all those $Q \in H_\infty^+$ such that

$$\|(T_{11}+T_{12}QT_{21})(s)\|_\infty \leqq \gamma \tag{1.2}$$

in which $T_{11}(s)$, $T_{12}(s)$, and $T_{21}(s)$ may always be chosen stable with $T_{12}(s)$ and $T_{21}(s)$ parts of inner matrices [10], [14], [15], [43], [46]. We change nothing by rewriting (1.2) as

$$\left\| \left(T_{11}+[T_\perp T_{12}]\begin{bmatrix} 0 & 0 \\ 0 & Q \end{bmatrix}\begin{bmatrix} \tilde{T}_\perp \\ T_{21} \end{bmatrix}\right)(s) \right\|_\infty \leqq \gamma, \qquad Q \in \mathscr{H}_\infty^+ \tag{1.3}$$

in which $T_\perp(s)$ and $\tilde{T}_\perp(s)$ are chosen to make $[T_\perp T_{12}](s)$ and $[\tilde{T}_\perp^\sim T_{21}^\sim]^\sim(s)$ inner. This, too, is always possible [2], [10], [16], [43]. Finally, by invoking the norm preserving property of inner matrices, we see that (1.3) is equivalent to the characterization of all $Q \in \mathscr{H}_\infty^{+,p\times m}$ such that

$$\left\| \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22}+Q \end{bmatrix}(s) \right\|_\infty \leqq \gamma \tag{1.4}$$

* Received by the editors May 16, 1988; accepted for publication (in revised form) January 26, 1990.

† Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, United Kingdom.

‡ Department of Electrical Engineering, Imperial College, Exhibition Road, London SW7 2BT, United Kingdom.

§ Department of Electrical Engineering, California Institute of Technology, Pasadena, California 91125.

¶ Department of Electrical Engineering Systems, University of Southern California, Los Angeles, California 90089-0781.

where

$$\begin{bmatrix} \dot{R}_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}(s) = \begin{bmatrix} T_{\perp}^{\sim} \\ T_{12} \end{bmatrix} T_{11}[\tilde{T}_{\perp}^{\sim} T_{21}^{\sim}](s) \in RH_{-}^{\infty}. \tag{1.5}$$

In this paper we will study the four block general distance problem given in (1.4) because of its intrinsic interest, and also due to its applicability to H^{∞} control.

Historical accounts of the development of solutions to this problem are contained in [11], [14] and [43]. Briefly, Doyle et al. [10], originally suggested a solution to the four block problem (1.4) based on the work of Davis et al. [9]. Other approaches included work on Hankel plus Toplitz operators in Jonckheere and Juang, [26] and on band extension problems in Dym and Gohberg [12]. These approaches certainly provide a theoretical solution, but when implemented on a computer, suffer from serious degree inflation problems. In an attempt to understand these inflation phenomena in the context of H^{∞} control, detailed cancellation analyses were carried out in [30], [31], [32], and [23] for cases of increasing complexity, and a controller with degree no greater than that of $P(s)$ in (1.1) was shown to exist. The outcome of this work also showed that solutions with $\deg(Q) \leqq \deg(R)$ exist to (1.4). These observations lead to the expectation that the original algorithm could be greatly improved and progress was made in [3] where just three Riccati equations of modest degree were required. The purpose of this paper is to present a new solution to the four block general distance problem requiring just two Riccati equations and also treating the optimal cases.

As indicated above, the solutions to the four block general distance problem can give representation formulae for all solutions to the H^{∞} control problem (1.2). One such formula is given in § 5 where the optimal cases are treated in detail. This solution requires two Riccati equations to be solved or, more precisely, (for certain optimal cases) for the appropriate stable invariant subspaces of two Hamiltonian matrices to be calculated. Solution formulae based on two Riccati equations were presented without proof in [20] for the suboptimal case and subsequently derivations for a limited class of plants were given in [11] using techniques similar to the state feedback results of Khargonekar, Petersen, and Rotea [28]. Formulae for certain optimal cases were given in [33], [35], and [45] using a descriptor representation and giving relations to interpolation. Many alternative and, in several cases, totally independent derivations of these results are now available using a variety of techniques. A solution based on J-spectral factorization theory is given in [22], while the related approach based on the notion of conjugation is employed in [29]. Hung has derived a formula in terms of two Riccati equations which deals with certain optimal situations [24], while Verma and Romig have a closed formula for one block problems [48]. An interesting by-product of this activity has been the discovery of a number of new interconnections. In [28], Khargonekar et al. note a connection between $\mathscr{H}^{\infty}$ control and game theory. The interplay between indefinite factorization and game theory, probably first noticed by Banker [4], has been rediscovered in the more general setting of $\mathscr{H}^{\infty}$ control [20], [22], [40]. The connection between risk sensitive optimal control [49] and game theory, originally discovered by Jacobson in the perfect information case [25], has also received renewed interest in the wider setting of $\mathscr{H}^{\infty}$ control [5], [20] and entropy minimization [21], [38]. Results on the finite horizon, time-varying case are given in [34] and [47], and finally, a solution applicable to distributed systems may be found in [13] and is due to Foias and Tannenbaum.

Section 2 contains a summary of the notation we will use. In § 3 we derive necessary and sufficient conditions for the existence of a suboptimal solution to the four block problem, and give a representation formula for all solutions. We treat the optimal case in detail in § 4. Section 5 deals with the application to $\mathcal{H}^\infty$ control theory. We derive necessary and sufficient conditions for the existence of a solution, and then give a formula for all solutions. By setting up the analysis in a descriptor framework, we are able to give a simple and complete treatment of all the optimal cases. In the event that $P_{11}(\infty)\neq 0$ and $P_{22}(\infty)\neq 0$, the controller formulae become cumbersome to write down. To obviate this difficulty, we employ the loop shifting transformations introduced in [45] to reduce the general case to a problem in which $P_{11}(\infty)=0$ and $P_{22}(\infty)=0$. We summarize the key findings of this work in the conclusions (§ 6).

2. Notation and Preliminaries.

The aim of this short section is to summarize the notation we intend to use; most of it is standard.

2.1. Notation.

$\mathbb{R}, \mathbb{R}_+, \mathbb{C}$	real, nonnegative, and complex numbers,
$\mathbb{R}(s)$	field of rational functions in s with real coefficients,
$\mathbb{C}_+, \mathbb{C}_-$	open right (respectively, left) half plane,
$\mathbb{F}^{p\times m}$	set of $p\times m$ matrices with elements in $\mathbb{F}(=\mathbb{R}, \mathbb{C}, \mathbb{R}(s)$ etc.),
$\lambda(A)$	spectrum of a square matrix A,
$\lambda_{\max}(A)$	eigenvalue of A with largest modulus,
A'	complex conjugate transpose of $A\in\mathbb{C}^{p\times m}$ (transpose if $A\in\mathbb{R}^{p\times m}$),
$A^{\#}$	generalized inverse,
$A^{\dagger}$	Moore-Penrose inverse,
$A\geqq 0, A>0$	A is positive semidefinite (respectively, positive definite),
$A\leqq 0, A<0$	A is negative semidefinite (respectively, negative definite),
$\mathcal{L}_\infty^{(p\times m)}$	space of $p\times m$ matrices with entries that are bounded on the $j\omega$-axis (including the point at ∞),
$\|\cdot\|_\infty$	$\mathcal{L}^\infty$-norm of matrices in $\mathcal{L}^\infty$,
$\mathcal{H}_\infty^{+,p\times m}$	subspace of $\mathcal{L}^\infty$; $p\times m$ matrices which are analytic and bounded in $\mathbb{C}_+$,
$\mathcal{H}_\infty^{-,p\times m}$	subspace of $\mathcal{L}^\infty$; $p\times m$ matrices which are analytic and bounded in $\mathbb{C}_-$,
$\|\cdot\|_H$	Hankel norm,
$\mathcal{RL}_\infty^{p\times m}$	same as $\mathcal{L}^{\infty(p\times m)}$ except elements are taken from $\mathbb{R}^{(p\times m)}(s)$,
$\mathcal{RH}_\infty^{+,p\times m}$	same as $\mathcal{H}_\infty^{+,(p\times m)}$ except elements are taken from $\mathbb{R}^{(p\times m)}(s)$,
$\mathcal{RH}_\infty^{-,p\times m}$	same as $\mathcal{H}_\infty^{-,(p\times m)}$ except elements are taken from $\mathbb{R}^{(p\times m)}(s)$,
$G^\sim(s)$	$G(-\bar{s})'$, the para-Hermitian conjugate of $G(s)$,
$\bar{\sigma}(A)=\|A\|_2$	the spectral norm of A,
Re s	real part of s,
$\mathcal{RH}_2(\mathcal{RH}_2^\perp)$	sets of functions $f(s)$ which are analytic in $\mathbb{C}_+(\mathbb{C}_-)$ such that $\sup_{\xi>0}\int_{-\infty}^{\infty}\|f(\xi+j\omega)\|_2^2\,d\omega<\infty(\sup_{\xi<0}\int_{-\infty}^{\infty}\|f(\xi+j\omega)\|_2^2\,d\omega<\infty)$,
$\rho(A)$	spectral radius.

Associated with a transfer function matrix $G(s)\in\mathbb{R}(s)^{p\times m}$ of McMillan degree $\leqq n$ is a state-space realization

$$G(s)=D+C(sI-A)^{-1}B \tag{2.1}$$

where $A\in\mathbb{C}^{n\times n}$, $B\in\mathbb{C}^{n\times m}$, $C\in\mathbb{C}^{p\times n}$, and $D\in\mathbb{C}^{p\times m}$. We will use the alternative notation $G(s)\overset{s}{=}(A, B, C, D)$ or

$$G(s)\overset{s}{=}\left[\begin{array}{c|c} A & B \\ \hline C & D \end{array}\right] \tag{2.2}$$

for realizations of $G(s)$. Generalized state-space models or descriptor system models [37] give rise to transfer functions via $G(s) = D + C(sE - A)^{-1}B \overset{s}{=} (A, B, C, D, E)$.

In the above notation, we have $G^{\sim}(s) \overset{s}{=} (-A', C', -B', D')$ and in the case that D is nonsingular, we have $G^{-1}(s) \overset{s}{=} (A - BD^{-1}C, BD^{-1}, -D^{-1}C, D^{-1})$. The system zeros of $G(s)$ are given by $\{\lambda(A - BD^{-1}C)\} \supseteq \{\text{McMillan zeros of } G(s)\}$; these sets are equal if the realization is minimal. If $G^{-1}(s) = G^{\sim}(s)$, then $G(s)$ is all-pass. $G(s)$ is called stable if all its poles are in $\mathbb{C}_-$.

We will talk about basis changes T in the state-space of $G(s)$; we will take this to mean $G(s) \overset{s}{=} (A, B, C, D) \xrightarrow{T} G(s) \overset{s}{=} (TAT^{-1}, TB, CT^{-1}, D)$. For descriptor system models, basis changes are given by

$$G(s) \overset{s}{=} (A, B, C, D, E) \xrightarrow{U,V} G(s) \overset{s}{=} (UAV, UB, CV, D, UEV).$$

We shall also make use of linear fractional transformations which are defined by

$$F_l\left\{\begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}, U\right\} = H_{11} + H_{12}U(I - H_{22}U)^{-1}H_{21} \tag{2.3}$$

where U is of dimension $l \times m$ if H_{22} has dimension $m \times l$.

3. Construction of an all-pass embedding. In this section we derive necessary and sufficient conditions for the existence of a $Q_{22}(s) \in \mathcal{RH}_\infty^+$ such that

$$\left\|\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} + Q_{22} \end{bmatrix}\right\|_\infty < 1 \tag{3.1}$$

for given $R_{ij}(s) \in \mathcal{RH}_\infty^{-,p_i \times m_j}$, and necessary conditions for the existence of $Q_{22}(s) \in \mathcal{RH}_\infty^+$ such that

$$\left\|\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} + Q_{22} \end{bmatrix}\right\|_\infty \leqq 1. \tag{3.2}$$

For simplicity, we will assume that

$$\bar{\sigma}([R_{11}(\infty), R_{12}(\infty)]) < 1 \tag{3.3}$$

and

$$\bar{\sigma}\left(\begin{bmatrix} R_{11}(\infty) \\ R_{21}(\infty) \end{bmatrix}\right) < 1. \tag{3.4}$$

If $\bar{\sigma}([R_{11}(j\omega), R_{12}(j\omega)]) < 1$ and $\bar{\sigma}([R_{11}^{\sim}(j\omega), R_{21}^{\sim}(j\omega)]) < 1$ for some ω, but not at $\omega = \infty$, then a bilinear transformation of the half plane into the half plane can give an equivalent problem satisfying (3.3) and (3.4). If however $\bar{\sigma}([R_{11}(j\omega), R_{12}(j\omega)]) = 1$ for all ω or $\bar{\sigma}([R_{11}^{\sim}(j\omega), R_{21}^{\sim}(j\omega)]) = 1$ for all ω, then significant modifications in detail need to be performed and so this case is not treated here in the interests of brevity and clarity.

Necessity of the conditions will be derived by assuming such a $Q_{22}(s)$ exists and finding an all-pass dilation of (3.2) to the following special structure, which is used to preserve the integrity of the first row and column of (3.1) and (3.2):

$$E_{aa}(s) = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 \\ p_4 \end{array} \begin{array}{c} \begin{array}{cccc} m_1 & m_2 & m_3 & m_4 \end{array} \\ \begin{bmatrix} R_{11} & R_{12} & R_{13} & 0 \\ R_{21} & E_{22} & E_{23} & E_{24} \\ R_{31} & E_{32} & E_{33} & E_{34} \\ 0 & E_{42} & E_{43} & E_{44} \end{bmatrix} \end{array} \tag{3.5}$$

where $E_{ij} = R_{ij} + Q_{ij}$, $i, j = 2, 3, 4$, $Q_{ij}(s) \in \mathcal{RH}_\infty^{+,p_i \times m_j}$, $R_{ij}(s) \in \mathcal{RH}_\infty^{-,p_i \times m_j}$. With the all-pass dilation constructed in a particular way (Lemma 3.1), so that R_{13}, R_{31}, $E_{24}^{\sim}$, $E_{42}^{\sim}$

being full rank in $\mathbb{C}_-$ (i.e., R_{13} and R_{31} are minimum phase), it turns out that (3.2) implies that (Proposition 3.2)

$$\|R_a^{\sim}\|_H \leqq 1 \tag{3.6}$$

where

$$R_a := \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \tag{3.7}$$

and R_a does not depend on Q_{22}. Similarly, it will be shown in Proposition 3.3 that (3.1) implies that

$$\|R_a^{\sim}\|_H < 1. \tag{3.8}$$

The sufficiency of (3.8) is derived in § 3.2 via a state-space construction along the lines of Glover [17], [18]. Finally, all solutions are derived from this particular solution of (3.1).

3.1. Necessary conditions. We now construct the all-pass dilation of (3.2).

LEMMA 3.1. *Suppose there exists* $Q_{22}(s) \in \mathcal{RH}_\infty^+$ *such that* (3.2) *holds with* R_{ij} *satisfying* (3.3) *and* (3.4). *Then there exists an all-pass dilation in the form* (3.5) *where*

$$E_{aa}E_{aa}^{\sim} = E_{aa}^{\sim}E_{aa} = I, \tag{3.9}$$

$$m_3 = p_1, \quad m_4 \leqq p_2, \quad p_3 = m_1, \quad p_4 \leqq m_2, \tag{3.10}$$

$$R_{13} \in \mathcal{RH}_\infty^-, \quad \text{rank } R_{13} = p_1 \quad \forall \text{ Re } s < 0, \tag{3.11}$$

$$R_{31} \in \mathcal{RH}_\infty^-, \quad \text{rank } R_{31} = m_1 \quad \forall \text{ Re } s < 0, \tag{3.12}$$

$$E_{23} := -[R_{21} \quad E_{22}][R_{11} \quad R_{12}]^{\sim}(R_{13}^{-1})^{\sim}, \tag{3.13}$$

$$E_{32} := -(R_{31}^{-1})^{\sim}[R_{11}^{\sim} \quad R_{21}^{\sim}]\begin{bmatrix} R_{12} \\ E_{22} \end{bmatrix}, \tag{3.14}$$

$$\text{rank } E_{24} = m_4 \quad \forall \text{ Re } s > 0, \tag{3.15}$$

$$\text{rank } E_{42} = p_4 \quad \forall \text{ Re } s > 0, \tag{3.16}$$

$$\begin{bmatrix} E_{33} & E_{34} \\ E_{43} & E_{44} \end{bmatrix} := -\begin{bmatrix} R_{31} & E_{32} \\ 0 & E_{42} \end{bmatrix}\begin{bmatrix} R_{11}^{\sim} & R_{21}^{\sim} \\ R_{12}^{\sim} & E_{22}^{\sim} \end{bmatrix}\begin{bmatrix} (R_{13}^{-1})^{\sim} & -(R_{13}^{-1})^{\sim}E_{23}^{\sim}(E_{24}^{l})^{\sim} \\ 0 & (E_{24}^{l})^{\sim} \end{bmatrix} \tag{3.17}$$

where $E_{24}^{l}E_{24} = I$, *with* E_{24}^{l} *analytic in* Re $s > 0$.

Proof. First let $R_{13}^{\sim}$ and $R_{31}^{\sim}$ be stable minimum phase spectral factors satisfying

$$R_{13}R_{13}^{\sim} = I - R_{11}R_{11}^{\sim} - R_{12}R_{12}^{\sim} \geqq 0 \quad \forall s = j\omega, \tag{3.18}$$

$$R_{31}^{\sim}R_{31} = I - R_{11}^{\sim}R_{11} - R_{21}^{\sim}R_{21} \geqq 0 \quad \forall s = j\omega \tag{3.19}$$

where the nonnegativity follows from (3.2). Furthermore, (3.3) implies rank $R_{13}(\infty) = p_1$ and (3.4) implies rank $R_{31}(\infty) = m_1$, hence R_{13} and R_{31} are square with R_{13}^{-1} and R_{31}^{-1} analytic in Re $s < 0$. Relations (3.2) and (3.9) also imply that

$$0 \leqq I - \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & E_{22} \end{bmatrix}\begin{bmatrix} R_{11}^{\sim} & R_{21}^{\sim} \\ R_{12}^{\sim} & E_{22}^{\sim} \end{bmatrix} = \begin{bmatrix} R_{13}R_{13}^{\sim} & R_{13}E_{23}^{\sim} \\ E_{23}R_{13}^{\sim} & I - R_{21}R_{21}^{\sim} - E_{22}E_{22}^{\sim} \end{bmatrix}$$

yielding

$$E_{23} = -(R_{21}R_{11}^{\sim} + E_{22}R_{12}^{\sim})(R_{13}^{\sim})^{-1}. \tag{3.20}$$

The second row of (3.5) may now be completed as follows:

(3.21)
$$E_{24}E_{24}^{\sim} := I - R_{21}R_{21}^{\sim} - E_{22}E_{22}^{\sim} - E_{23}E_{23}^{\sim} \geqq 0.$$

Similarly,

(3.22)
$$\begin{aligned} E_{32} &= -(R_{31}^{\sim})^{-1}(R_{11}^{\sim}R_{12} + R_{21}^{\sim}E_{22}), \\ E_{42}^{\sim}E_{42} &:= I - R_{12}^{\sim}R_{12} - E_{22}^{\sim}E_{22} - E_{23}^{\sim}E_{23} \geqq 0. \end{aligned}$$

It is then simply verified that (3.17) completes the dilation. □

Now let us examine the partially agumented system

(3.23)
$$E_a := \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & E_{22} & E_{23} \\ R_{31} & E_{32} & E_{33} \end{bmatrix}$$

in more detail. First, note that since E_{aa} is all-pass by Lemma 3.1, that

(3.24)
$$\|E_a\|_\infty \leqq 1.$$

Also by (3.17) and (3.9)

$$\begin{aligned} E_{33} &= -[R_{31} \quad E_{32}][R_{11} \quad R_{12}]^{\sim}(R_{13}^{-1})^{\sim} \\ &= -(R_{31}^{-1})^{\sim}[I - R_{11}^{\sim}R_{11} - R_{21}^{\sim}R_{21} \quad -R_{11}^{\sim}R_{12} - R_{21}^{\sim}E_{22}][R_{11} \quad R_{12}]^{\sim}(R_{13}^{-1})^{\sim} \end{aligned}$$

(3.25)
$$= -(R_{31}^{-1})^{\sim}R_{11}^{\sim}(R_{13}^{-1})^{\sim} + (R_{31}^{-1})^{\sim}[R_{11}^{\sim} \quad R_{21}^{\sim}]\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & E_{22} \end{bmatrix}\begin{bmatrix} R_{11}^{\sim} \\ R_{12}^{\sim} \end{bmatrix}(R_{13}^{-1})^{\sim}.$$

Hence,

(3.26)
$$E_a = T_1\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & E_{22} \end{bmatrix}T_2 + T_3$$

where

(3.27)
$$T_1 = \begin{bmatrix} I & 0 \\ 0 & I \\ -(R_{31}^{-1})^{\sim}R_{11}^{\sim} & -(R_{31}^{-1})^{\sim}R_{21}^{\sim} \end{bmatrix},$$

(3.28)
$$T_2 = \begin{bmatrix} I & 0 & -R_{11}^{\sim}(R_{13}^{-1})^{\sim} \\ 0 & I & -R_{12}^{\sim}(R_{13}^{-1})^{\sim} \end{bmatrix},$$

(3.29)
$$T_3 = \begin{bmatrix} 0 & 0 & (R_{13}^{-1})^{\sim} \\ 0 & 0 & 0 \\ (R_{31}^{-1})^{\sim} & 0 & -(R_{31}^{-1})^{\sim}R_{11}^{\sim}(R_{13}^{-1})^{\sim} \end{bmatrix}.$$

(3.30) Note that T_1, T_2, and T_3 are analytic in Re $s > 0$.

Relation (3.26) implies that

(3.31)
$$E_a = T_1\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}T_2 + T_1\begin{bmatrix} 0 & 0 \\ 0 & Q_{22} \end{bmatrix}T_2 + T_3.$$

Now define R_{23}, R_{32}, and R_{33} such

(3.32)
$$\begin{aligned} R_a &:= \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \\ &= (\text{constant}) + \left\{T_1\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}T_2\right\}_{\text{anticausal}}. \end{aligned}$$

That is,

$$E_a = R_a + T_4Q_{22}T_5 + T_6 \tag{3.33}$$

where $T_4 = T_1[0 \quad I]'$; $T_5 = [0 \quad I]T_2$; $(T_4Q_{22}T_5 + T_6) \in \mathcal{RH}_\infty^+$ since T_4, T_5, T_6, Q_{22} are analytic in $\operatorname{Re}(s) > 0$ and $\|E_a\|_\infty \leqq 1$ implies that $(T_4Q_{22}T_5 + T_6)$ has no poles on the imaginary axis. Hence (3.24), (3.33) and Nehari's theorem give that $\|R_a^{\sim}\|_H \leqq 1$.

The important observation here is that R_a can be constructed without explicit knowledge of Q_{22}. Moreover, R_{11}, R_{12}, R_{13}, R_{21}, R_{22}, R_{31} uniquely determine R_a. The following proposition has thus been established.

PROPOSITION 3.2. *If there exists* $Q_{22} \in \mathcal{RH}_\infty^+$ *such that* (3.2) *holds with* R_{ij} *satisfying* (3.3) *and* (3.8), *then*

$$\|R_{11} \quad R_{12}\|_\infty \leqq 1, \tag{3.34}$$

$$\left\| \begin{matrix} R_{11} \\ R_{21} \end{matrix} \right\|_\infty \leqq 1, \tag{3.35}$$

and

$$\|R_a^{\sim}\|_H \leqq 1 \tag{3.36}$$

where R_a *is given in* (3.32). □

For the case of strict inequality in (3.1), all the above inequalities can be made strict and this is now proven.

PROPOSITION 3.3. *If there exists* $Q_{22} \in \mathcal{RH}_\infty^+$ *such that* (3.1) *holds with* R_{ij} *satisfying* (3.3) *and* (3.4), *then*

$$\|R_{11} \quad R_{12}\|_\infty < 1, \tag{3.37}$$

$$\left\| \begin{matrix} R_{11} \\ R_{21} \end{matrix} \right\|_\infty < 1, \tag{3.38}$$

and

$$\|R_a^{\sim}\|_H < 1 \tag{3.39}$$

where R_a *is given in* (3.32). □

Proof. Formulae (3.37) and (3.38) are immediate consequences of (3.1) and also imply that R_{13}, R_{31} may be chosen with inverses in $\mathcal{RH}_\infty^-$. Note that in the expression for E_a given in (3.33), R_a, T_4, T_5, and T_3 are all independent of Q_{22}. Furthermore, for $\Delta \in \mathcal{RH}_\infty^+$, define

$$E_\alpha(\Delta) := R_a + T_4(Q_{22} + \Delta)T_5 + T_6. \tag{3.40}$$

Then $\|E_a(\Delta)\|_\infty \leqq 1$ for all Δ such that

$$\|\Delta\|_\infty < 1 - \left\| \begin{matrix} R_{11} & R_{12} \\ R_{21} & E_{22} \end{matrix} \right\|_\infty. \tag{3.41}$$

Since the construction of Lemma 3.1 will still work, we will now suppose that $\|R_a^{\sim}\|_H = 1$ and construct a contradiction. For all Δ satisfying (3.41), we have

$$E_a(\Delta)V(s) = U(-s) \tag{3.42}$$

where V and U are the Laplace transforms of the corresponding Schmidt vectors for the Hankel operator corresponding to $R_a^{\sim}$ [14], [36]. Note that $U, V \in \mathcal{RH}_2$. Hence $(E_a(\Delta) - E_a(0))V(s) = 0 \Rightarrow T_4\Delta T_5V(s) = 0$ for all Δ satisfying (3.41). Thus

$$T_5V(s) = 0. \tag{3.43}$$

Now consider the last block row of (3.42)

$$[R_{31} \quad E_{32} \quad E_{33}]V(s) = [0 \quad 0 \quad I]U(-s)$$

and substitute for E_{33} as

$$E_{33} = -(R_{31}R_{11}^{\sim} + E_{32}R_{12}^{\sim})(R_{13}^{-1})^{\sim}$$

yielding

$$R_{31}[I \quad 0 \quad -R_{12}^{\sim}(R_{13}^{-1})^{\sim}]V(s) + E_{32}[0 \quad I \quad -R_{12}^{\sim}(R_{13}^{-1})]V(s) = [0 \quad 0 \quad I]U(-s).$$

The second term is $E_{32}T_5V(s) = 0$ by (3.43). Hence

$$[I \quad 0 \quad -R_{11}^{\sim}(R_{13}^{-1})^{\sim}]V(s) = R_{31}^{-1}[0 \quad 0 \quad I]U(-s). \tag{3.44}$$

The left-hand side of (3.44) is in $\mathscr{RH}_2$ where the right-hand side is in $\mathscr{RH}_2^{\perp}$ and hence both must be zero. This together with (3.28) and (3.43) gives that $T_2V(s) = 0$ which, when substituted into (3.42) using (3.26), gives

$$U(-s) = T_3V(s)$$

but $U(-s) \in \mathscr{RH}_2^{\perp}$ and $T_3V(s) \in \mathscr{RH}_2$ so that both must be zero contradicting U being the Laplace transform of a Schmidt vector. □

3.2. State-space construction and sufficient conditions. We will first construct a state-space description of $\mathrm{R_a}$ given in (3.32). The terms R_{13} and R_{31} come from standard spectral factorization problems and it is a routine exercise to find out the realization of R_a.

LEMMA 3.4. *Let*

$$\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \in \mathscr{RH}_{\infty}^{-}$$

be such that (3.2) *holds for some* $Q_{22} \in \mathscr{RL}_{\infty}$ *and have the state-space realization*

$$\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \overset{s}{=} \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & 0 \end{array}\right] \tag{3.45}$$

where $\operatorname{Re} \lambda_i(A) > 0$ *for all i. Then* R_a *given by* (3.32) *has a state-space realization of the form*

$$R_a \overset{s}{=} \left[\begin{array}{c|ccc} A & B_1 & B_2 & B_3 \\ \hline C_1 & D_{11} & D_{12} & D_{13} \\ C_2 & D_{21} & * & * \\ C_3 & D_{31} & * & * \end{array}\right] \tag{3.46}$$

where

$$D_{13}D_{13}' = I - D_{11}D_{11}' - D_{12}D_{12}' > 0, \tag{3.47}$$

$$D_{31}'D_{31} = I - D_{11}'D_{11} - D_{21}'D_{21} > 0, \tag{3.48}$$

$$B_3 = (XC_1' - [B_1 \quad B_2][D_{11} \quad D_{12}]')(D_{13}^{-1})', \tag{3.49}$$

$$C_3 = (D_{31}^{-1})'(B_1'Y - [D_{11}' \quad D_{21}'][C_1' \quad C_2']'). \tag{3.50}$$

$X = X' \geqq 0$ *is the unique solution to the algebraic Riccati equation*

$$-XA' - AX + [B_1 \quad B_2][B_1 \quad B_2]' + B_3B_3' = 0 \tag{3.51}$$

such that $\operatorname{Re}\lambda_i(A-B_3D_{13}^{-1}C_1)\geqq 0$ *for all i.* $Y=Y'\geqq 0$ *is the unique solution to the algebraic Riccati equation*

$$-YA-A'Y+[C_1' \quad C_2'][C_1' \quad C_2']'+C_3'C_3=0 \tag{3.52}$$

such that $\operatorname{Re}\lambda_i(A-B_1D_{31}^{-1}C_3)\geqq 0$ *for all i.*

Proof. The construction of B_3 and D_{13} to form R_{13}, and C_3 and D_{31} to form R_{31} uses standard techniques for calculating spectral factors using Riccati equations [1], [50].

The realization of R_a clearly matches R_{11}, R_{12}, R_{13}, R_{21}, R_{22}, and R_{31}. It remains to verify that R_{23}, R_{32}, and R_{33} given in (3.32) match the given realization. This is a routine state-space manipulation and uses the following realizations:

$$(R_{31}^{-1})^{\sim}[R_{11}^{\sim} \quad R_{21}^{\sim}] \overset{s}{=} \left[\begin{array}{c|c} -A'+C_3'(D_{31}^{-1})'B_1' & C_3'(D_{31}^{-1})'[D_{11}' \quad D_{21}']-[C_1' \quad C_2'] \\ \hline (D_{31}^{-1})'B_1' & (D_{31}^{-1})'[D_{11}' \quad D_{21}'] \end{array}\right]$$

and the dual

$$\begin{bmatrix} R_{11}^{\sim} \\ R_{12}^{\sim} \end{bmatrix}(R_{13}^{-1})^{\sim} \overset{s}{=} \left[\begin{array}{c|c} -A'+C_1'(D_{13}^{-1})'B_3' & C_1'(D_{13}^{-1})' \\ \hline \begin{bmatrix} D_{11}' \\ D_{12}' \end{bmatrix}(D_{13}^{-1})'B_3'-\begin{bmatrix} B_1' \\ B_2' \end{bmatrix} & \begin{bmatrix} D_{11}' \\ D_{12}' \end{bmatrix}(D_{13}^{-1})' \end{array}\right].$$

Equations (3.50) and (3.52) then give that

(3.53)

$$(R_{31}^{-1})^{\sim}[R_{11}^{\sim} \quad R_{21}^{\sim}]\begin{bmatrix} C_1 \\ C_2 \end{bmatrix}(sI-A)^{-1}=-C_3(sI-A)^{-1}+\{\text{terms in analytic in } \operatorname{Re}(s)>0\}$$

and (3.49) and (3.51) imply that

(3.54)

$$(sI-A)^{-1}[B_1 \quad B_2]\begin{bmatrix} R_{11}^{\sim} \\ R_{21}^{\sim} \end{bmatrix}(R_{13}^{-1})^{\sim}=-(sI-A)^{-1}B_3+\{\text{terms in analytic in } \operatorname{Re}(s)>0\}.$$

The definition of R_a in (3.32) together with (3.53) and (3.54) then give the result. □

We now immediately have the following corollaries of Propositions 3.2 and 3.3, on noting that X and Y are the Gramians for R_a and hence $\|R_a\|_H^2=\rho(XY)$ [14], [6].

COROLLARY 3.5. *Let*

$$\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}\in \mathcal{RH}_\infty^-$$

satisfy (3.3) *and* (3.4) *with the state-space realization of* (3.45). *Then*

(i) *If there exists* $Q_{22}\in\mathcal{RH}_\infty^+$ *satisfying* (3.2), *then* $\rho(XY)\leqq 1$.

(ii) *If there exists* $Q_{22}\in\mathcal{RH}_\infty^+$ *satisfying* (3.1), *then* $\rho(XY)<1$. □

Now that we have an explicit state-space realization of R_a and a condition on the Gramians we can attempt a state-space constrution of an all-pass embedding along the lines of (3.3) (without assuming knowledge of a candidate Q_{22}). This construction can be carried out along the lines of Glover [17], [18] and will then give a sufficiency proof. We will consider the case $\|R_a\|_H<1$ in this section with the $\|R_a\|_H=1$ considered in § 4.

First, consider the unitary dilation of the D-matrix. Such a dilation can be constructed in the form

$$(3.55)\qquad D_e = \begin{array}{c} \\ p_1 \\ p_2 \\ p_3 = m_1 \\ m_2 \geqq p_4 \end{array} \begin{array}{c} \begin{array}{cccc} m_1 & m_2 & m_3 = p_1 & m_4 \leqq p_2 \end{array} \\ \left[\begin{array}{cccc} D_{11} & D_{12} & D_{13} & 0 \\ D_{21} & D_{22} & D_{23} & D_{24} \\ D_{31} & D_{32} & D_{33} & D_{34} \\ 0 & D_{42} & D_{43} & 0 \end{array}\right] \end{array}.$$

In (3.55) $\left[\begin{smallmatrix} D_{24} \\ D_{34} \end{smallmatrix}\right]$ is chosen to be an orthonormal basis for the nullspace of $[D_{21}' \quad D_{31}']$ (which has dimension p_2 since rank $D_{31} = m_1$, and also implies that D_{24}^{-1} exists). Similarly, $[D_{42} \quad D_{43}]'$ is an orthonormal basis for the nullspace of $[D_{12} \quad D_{13}]$, and D_{42}^{-1} exists. The remaining terms are then uniquely given by

$$(3.56)\qquad \begin{bmatrix} D_{22} & D_{23} \\ D_{32} & D_{33} \end{bmatrix} = -\begin{bmatrix} D_{21}' & D_{31}' \\ D_{24}' & D_{34}' \end{bmatrix}^{-1} \begin{bmatrix} D_{11}' & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} D_{12} & D_{13} \\ D_{42} & D_{43} \end{bmatrix}.$$

Now suppose that we can find a realization of $E_{aa} = R_{aa} + Q_{aa}$ in the form

$$(3.57)\qquad R_{aa} \overset{s}{=} \left[\begin{array}{c|c} A & B_e \\ \hline C_e & * \end{array}\right] \overset{s}{=} \begin{bmatrix} R_a & 0 \\ 0 & 0 \end{bmatrix},$$

$$(3.58)\qquad Q_{aa} \overset{s}{=} \left[\begin{array}{c|c} \hat{A} & \hat{B}_e \\ \hline \hat{C}_e & 0 \end{array}\right], \quad \operatorname{Re} \lambda_i(\hat{A}) < 0 \quad \forall i$$

where

$$(3.59)\qquad B_e = [B_1 \quad B_2 \quad B_3 \quad 0],$$

$$(3.60)\qquad C_e' = [C_1' \quad C_2' \quad C_3' \quad 0],$$

$$(3.61)\qquad \hat{B}_e = [0 \quad \hat{B}_2 \quad \hat{B}_3 \quad \hat{B}_4],$$

$$(3.62)\qquad \hat{C}_e' = [0 \quad \hat{C}_2' \quad \hat{C}_3' \quad \hat{C}_4'],$$

and

$$(3.63)\qquad E_{aa} \overset{s}{=} \left[\begin{array}{cc|c} A & 0 & B_e \\ 0 & \hat{A} & \hat{B}_e \\ \hline C_e & \hat{C}_e & D_e \end{array}\right].$$

In order to construct $\hat{A}$, $\hat{B}_e$, and $\hat{C}_e$ such that E_{aa} is all-pass, we will write the all-pass lemma equations of Glover [16, Thm. 5.1] with postulated solutions to the Lyapunov equations as follows:

$$(3.64)\qquad X_e = \begin{bmatrix} -X & I \\ I & YZ^{-1} \end{bmatrix}$$

where

$$(3.65)\qquad Z = I - XY,$$

$$(3.66)\qquad Y_e = X_e^{-1} = \begin{bmatrix} -Y & Z' \\ Z & ZX \end{bmatrix}.$$

(The form of X_e and Y_e essentially comes from Glover [16, Lemma 8.2] but this need not concern us.) With this value for X_e, $\hat{A}$, $\hat{B}_e$, $\hat{C}_e$ can be constructed to satisfy the conditions of [16, Thm. 5.1], viz. D_e unitary and

$$\begin{bmatrix} A & 0 \\ 0 & \hat{A} \end{bmatrix} X_e + X_e \begin{bmatrix} A' & 0 \\ 0 & \hat{A}' \end{bmatrix} + \begin{bmatrix} B_e \\ \hat{B}_e \end{bmatrix} [B_e' \quad \hat{B}_e'] = 0, \tag{3.67}$$

$$D_e[B_e' \quad \hat{B}_e'] + [C_e \quad \hat{C}_e] X_e = 0. \tag{3.68}$$

Postmutiplying (3.68) by the invertible matrix $\left[\begin{smallmatrix} I & -Y \\ 0 & Z \end{smallmatrix}\right]$ yields the equivalent expressions

$$\hat{C}_e = C_e X - D_e B_e' \tag{3.69}$$

and

$$-D_e B_e' Y + D_e \hat{B}_e' Z + C_e = 0.$$

Since Z is nonsingular

$$\hat{B}_e = (Z^{-1})'(Y B_e - C_e' D_e) \tag{3.70}$$

and (3.68) holds.

Now let E be the left-hand side of (3.67); then $[I \quad 0]E[I \quad 0]' = 0$ by (3.51) and $[I \quad 0]Y_e E Y_e [I \quad 0]' = 0$ by (3.52). $\hat{A}$ will be chosen to make $[I \quad 0]E[0 \quad I]' = 0$ as

$$\hat{A} = -A' - \hat{B}_e B_e' \tag{3.71}$$

$$= (Z^{-1})'\{-(I - YX)A' - (YB_e - C_e' D_e) B_e'\}$$

$$= (Z^{-1})'\{-A' - YAX + C_e' D_e B_e'\}. \tag{3.72}$$

With this value of $\hat{A}$ we obtain $[I \quad 0]E = 0$ and hence

$$\begin{bmatrix} I & 0 \\ -Y & Z' \end{bmatrix} E \begin{bmatrix} I & -Y \\ 0 & Z \end{bmatrix} = 0$$

and thus $E = 0$. Therefore, with these values of $\hat{A}$, $\hat{B}_e$, and $\hat{C}_e$, E_a satisfies (3.67) and (3.68) and is hence all-pass. Furthermore, $YZ^{-1} \geqq 0$ since $\rho(XY) < 1$ by Corollary 3.5 and hence $\operatorname{Re} \lambda_i(\hat{A}) < 0$ (by Wonham [52, Lemma 12.2, p. 227]) since $\hat{A} Y Z^{-1} + Z'^{-1} Y \hat{A}' + \hat{B}_e \hat{B}_e' = 0$ and $(\hat{A}, \hat{B}_e)$ is clearly stabilizable. Therefore $Q_{aa} \in \mathcal{RH}_\infty^+$ and it remains to show that $\hat{B}_e$ and $\hat{C}_e$ have the zero terms given in (3.61) and (3.62). From (3.70) and (3.50)

$$Z' \hat{B}_e [I \quad 0 \quad 0 \quad 0]' = Y B_1 - [C_1' \quad C_2' \quad C_3'] \begin{bmatrix} D_{11} \\ D_{21} \\ D_{31} \end{bmatrix} = 0.$$

Similarly, (3.69) and (3.49) give that

$$[I \quad 0 \quad 0 \quad 0] \hat{C}_e = C_1 X - [D_{11} \quad D_{12} \quad D_{13}] \begin{bmatrix} B_1' \\ B_2' \\ B_3' \end{bmatrix} = 0.$$

Therefore we have verified that Q_{aa} given by (3.58), (3.72), (3.69), and (3.70) satisfy (3.61) and (3.62) and, for all unitary D_e, gives an all-pass E_{aa} with $Q_{aa} \in \mathcal{RH}_\infty^+$. Note that once the form of D_e and X_e have been specified, all the other terms are uniquely determined and the required zero structure on Q_{aa} has been ensured by fixing the form of the first row and column of R_{aa} to be as in (3.5). All solutions can also be generated from this Q_{aa} as will now be stated in the main result of this section.

THEOREM 3.6. *Let $R_{ij} \in \mathcal{RH}_\infty^{-,p_i \times m_j}$ for $i, j = 1, 2$ have the realization*

$$\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \overset{s}{=} \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & * \end{array}\right]$$

with $\operatorname{Re} \lambda_i(A) > 0$ *for all i, and*

$$\bar{\sigma}[D_{11} \quad D_{12}] < 1, \qquad \bar{\sigma}[D'_{11} \quad D'_{21}] < 1.$$

(a) *Then there exists $Q \in \mathcal{RH}_\infty^+$ such that*

$$\left\| \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} + Q \end{bmatrix} \right\|_\infty < 1 \tag{3.73}$$

if and only if (3.37), (3.38) *and* $\|R_a\|_H < 1$, *or equivalently if and only if* (3.37), (3.38) *and* $\rho(XY) < 1$ *where* $X, Y \geqq 0$ *are as defined in Lemma* 3.4.

(b) *If the conditions of part* (a) *are satisfied, then each* $Q \in \mathcal{RH}_\infty^+$ *satisfying* (3.73) *is given by*

$$Q = F_l \left\{ \begin{bmatrix} Q_{22} & Q_{24} \\ Q_{42} & Q_{44} \end{bmatrix}; \Phi \right\} \tag{3.74}$$

for some $\Phi \in \mathcal{RH}_\infty^+$ *and* $\|\Phi\|_\infty < 1$, *where*

$$\begin{bmatrix} Q_{22} & Q_{24} \\ Q_{42} & Q_{44} \end{bmatrix} \overset{s}{=} \left[\begin{array}{c|cc} \hat{A} & \hat{B}_2 & \hat{B}_4 \\ \hline \hat{C}_2 & D_{22} & D_{24} \\ \hat{C}_4 & D_{42} & 0 \end{array}\right] \tag{3.75}$$

with D_{ij}, D_e *given by* (3.55). *The remaining matrices are given by*

$$\begin{bmatrix} \hat{C}_2 \\ \hat{C}_4 \end{bmatrix} = \begin{bmatrix} C_2 X \\ 0 \end{bmatrix} - \begin{bmatrix} D_{21} & D_{22} & D_{23} \\ 0 & D_{42} & D_{43} \end{bmatrix} \begin{bmatrix} B'_1 \\ B'_2 \\ B'_3 \end{bmatrix}, \tag{3.76}$$

$$[\hat{B}_2 \quad \hat{B}_4] = (Z^{-1})' \{ Y[B_2 \quad 0] - [C'_1 \quad C'_2 \quad C'_3] \} \begin{bmatrix} D_{12} & 0 \\ D_{22} & D_{24} \\ D_{32} & D_{34} \end{bmatrix}, \tag{3.77}$$

$$Z = I - XY, \qquad \hat{A} = (Z^{-1})' \{ -A' - YAX + C'_e D_e B'_e \}. \tag{3.78}$$

(c) *If R_{ij} satisfies* (3.34) *and* (3.35) *and* $\rho(XY) < 1$, *then every solution* $Q \in \mathcal{RH}_\infty^+$ *such that*

$$\left\| \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} + Q \end{bmatrix} \right\|_\infty \leqq 1$$

is given by (3.74) *for some* $\Phi \in \mathcal{RH}_\infty^+$ *with* $\|\Phi\|_\infty \leqq 1$.

Proof. Part (a) follows from Propositions 3.2 and 3.3, Corollary 3.5, and the construction preceding the theorem statement. Part (b) follows in a similar way to that given in Glover [17], [18] as follows. First, recall that E_{aa} is all-pass where

$$E_{aa}(s) = \begin{bmatrix} R_{11} & R_{12} & R_{13} & 0 \\ R_{21} & E_{22} & E_{23} & Q_{24} \\ R_{31} & E_{32} & E_{33} & Q_{34} \\ 0 & Q_{42} & Q_{43} & Q_{44} \end{bmatrix}$$

and hence,

$$F_l\left\{E_{aa}, \begin{bmatrix} 0 & 0 \\ 0 & \Phi \end{bmatrix}\right\} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22}+Q \end{bmatrix} \quad \text{satisfies} \quad \left\| \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22}+Q \end{bmatrix} \right\|_\infty < 1$$

if $\|\Phi\|_\infty < 1$ [19], [42] where

$$Q = F_l\left\{\begin{bmatrix} Q_{22} & Q_{24} \\ Q_{42} & Q_{44} \end{bmatrix}; \Phi\right\} \in \mathcal{RH}_\infty^+.$$

That Q is in $\mathcal{RH}_\infty^+$ follows from a small gain argument since $Q_{ij} \in \mathcal{RH}_\infty^+$ and $\Phi \in \mathcal{RH}_\infty^+$ with $\|Q_{44}\Phi\|_\infty < 1$. Hence all such Q satisfy (3.73).

To show that all such Q can be expressed as in (3.74) we first show that Q_{24}^{-1}, $Q_{42}^{-1} \in \mathcal{RH}_\infty^+$. The A-matrix of Q_{24}^{-1} is given by

$$\begin{aligned} \hat{A} - \hat{B}_2 D_{42}^{-1} \hat{C}_4 &= -A' - \hat{B}_e B_e' - \hat{B}_2(-B_2' - D_{42}^{-1} D_{43} B_3') \\ &= -A' - (\hat{B}_3 - \hat{B}_2 D_{42}^{-1} D_{43}) B_3' \end{aligned}$$

since D_e is unitary, $D_{42}^{-1} D_{43} = -D_{12}'(D_{13}^{-1})'$; hence,

$$\begin{aligned} \hat{B}_3 - \hat{B}_2 D_{42}^{-1} D_{43} &= [\hat{B}_2 \quad \hat{B}_3] \begin{bmatrix} D_{12}' \\ D_{13} \end{bmatrix} (D_{13}^{-1})' \\ &= \hat{B}_e D_e' \begin{bmatrix} I \\ 0 \\ 0 \\ 0 \end{bmatrix} (D_{13}^{-1})' \quad \text{by (3.70)} \\ &= (Z^{-1})'\left(Y B_e D_e' \begin{bmatrix} I \\ 0 \\ 0 \\ 0 \end{bmatrix} - C_1' \right)(D_{13}^{-1})' \\ &= (Z^{-1})'(YX - I)C_1'(D_{13}^{-1})' \quad \text{by (3.49)} \\ &= -C_1'(D_{13}^{-1})'. \end{aligned}$$

Hence $\hat{A} - \hat{B}_2 D_{42}^{-1} \hat{C}_4 = -(A' - C_1'(D_{13}^{-1})' B_3')$ and $Q_{42}^{-1} \in \mathcal{RH}_\infty^+$ since $\operatorname{Re} \lambda_i(A - B_3 D_{13}^{-1} C_1) < 0$ are the zeros of R_{13} which, by construction, has no zeros in $\operatorname{Re} s > 0$. See (3.51).

A similar argument gives that $\hat{A} - \hat{B}_4 D_{24}^{-1} \hat{C}_2 = -(Z^{-1})'(A - B_1 D_{31}^{-1} C_3)' Z'$ and hence $Q_{24}^{-1} \in \mathcal{RH}_\infty^+$. Suppose that Q satisfies (3.73) and define Φ such that (3.74) holds, that is,

$$\Psi := Q_{24}^{-1}(Q - Q_{22}) Q_{42}^{-1} = \Phi(I - Q_{44}\Phi)^{-1}$$

and

$$\Phi = -(I + \Psi Q_{44})^{-1} \Psi$$

and Φ clearly exists as a proper rational function in $\mathcal{L}_\infty$. Furthermore $\|\Phi\|_\infty < 1$ since

$$\left\| F_l\left(E_{aa}, \begin{pmatrix} 0 & 0 \\ 0 & \Phi \end{pmatrix}\right) \right\|_\infty < 1, \qquad Q_{44}(\infty) = 0$$

(see [30]). Suppose that Φ has a coprime factorization over $\mathcal{RH}_\infty^+$ as $\Phi = UV^{-1}$ with $U, V \in \mathcal{RH}_\infty^+$. Since U, V is coprime and Q_{44} is stable, U, $(V - Q_{44}U)$ is also coprime, and a coprime factorization of $\Psi = U(V - Q_{44}U)^{-1}$. Since $\Psi \in \mathcal{RH}_\infty^+$ the winding number of $\det(V - Q_{44}U)(j\omega)$ around the origin must be zero. However, $\det(V - Q_{44}U)(j\omega) = \det V(j\omega) \det(I - Q_{44}\Phi)(j\omega)$ and hence the winding number of $\det V(j\omega)$ is zero, since that of $\det(I - Q_{44}\Phi)(j\omega)$ is zero (because $\|Q_{44}\Phi\|_\infty < 1$) hence $\Phi \in \mathcal{RH}_\infty^+$.

(c) The major difference between this case and part (b) is that R_{13}^{-1}, R_{31}^{-1} need not be in $\mathcal{RH}_\infty^-$ and Q_{24}^{-1}, Q_{42}^{-1} need not be in $\mathcal{RH}_\infty^+$; however, R_{13}, R_{31}, Q_{24}, and Q_{42} will still be full rank in the appropriate halfplanes and only rank deficient at a finite number of points on $s = j\omega$. The proof that (3.74) with $\|\Phi\|_\infty \leqq 1$ gives a class of solutions for Q is the same as in part (b) except that we now need to prove that $\|Q_{44}\|_\infty < 1$. Suppose $\|Q_{44}\|_\infty = 1$. Since E_{aa} is all-pass if $\bar{\sigma}(Q_{44}(j\omega_0)) = 1$ for some ω_0 then $\operatorname{rank}[Q_{42}(j\omega_0) \quad Q_{43}(j\omega_0)] < p_4 = m_3 = p_1$, i.e., $[Q_{42} \quad Q_{43}]$ has an imaginary axis zero. Now

$$[Q_{42} \quad Q_{43}] \overset{s}{=} \left[\begin{array}{c|cc} \hat{A} & \hat{B}_2 & \hat{B}_3 \\ \hline \hat{C}_4 & D_{42} & D_{43} \end{array}\right]$$

$$= [D_{42} \quad D_{43}] \left[\begin{array}{c|cc} \hat{A} & \hat{B}_2 & \hat{B}_3 \\ \hline -B_2' & I & 0 \\ -B_3' & 0 & I \end{array}\right] \quad \text{from (3.76)}$$

and

$$\hat{A} + [\hat{B}_2 \quad \hat{B}_3][B_2 \quad B_3]' = -A' \quad \text{by (3.71)}$$

and hence the zeros of $[Q_{42} \quad Q_{43}]$ can only appear at $\lambda_i(-A')$. But $\operatorname{Re} \lambda_i(-A') < 0$. Thus $\operatorname{rank}[Q_{42}(j\omega) Q_{43}(j\omega)] = p_4 = m_3 = p_1$ for all ω, giving $\|Q_{44}\|_\infty < 1$.

To prove that (3.74) with $\|\Phi\|_\infty \leqq 1$ gives all solutions we first need that

$$\left\| F_l\left(E_{aa}, \begin{pmatrix} 0 & 0 \\ 0 & \Phi \end{pmatrix}\right) \right\|_\infty \leqq 1$$

implies $\|\Phi\|_\infty \leqq 1$, when R_{13}, R_{31}, Q_{24}, Q_{42} are only full rank for almost all $s = j\omega$. This is simply proved by noting that if $\bar{\sigma}(\Phi(j\omega_0)) > 1$ then $\bar{\sigma}(\Phi(s)) > 1$ for all s in some neighbourhood of $j\omega_0$, and hence we can find ω_1 such that $\bar{\sigma}(\Phi(j\omega_1)) > 1$ and R_{13}, R_{31}, E_{24}, and E_{42} all have full rank at $j\omega_1$; a contradiction is then easily established.

Finally, defining Ψ and Φ as in (b) with $\Phi = UV^{-1}$ we obtain that $\det V(j\omega) \neq 0$ for all ω since $\|\Phi\|_\infty \leqq 1$, $\det(I - Q_{44}\Phi)(j\omega) \neq 0$ for all ω since $\|Q_{44}\Phi\|_\infty < 1$, and hence $\Psi = U(V - Q_{44}U)^{-1}$ and $\det(V - Q_{44}U) = \det(V) \det(I - Q_{44}\Phi) \neq 0$ for all ω implies that $\Psi \in \mathcal{RH}_\infty$ and the result follows as in (b). □

Remark. Theorem 3.6 gives a complete solution to the problem when $\rho(XY) < 1$, the case when $\rho(XY) = 1$ is substantially more involved and is given in § 4. The solution of

$$\left\| \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} + Q \end{bmatrix} \right\|_\infty \leqq \gamma$$

for the minimum possible γ will require an iterative search on γ with the problem scaled to γ at each step (e.g., scale $D_{ij} \to \gamma^{-1} D_{ij}$, $B_i \to \gamma^{-1/2} B_i$, $C_i \to \gamma^{-1/2} C_i$). A value of γ will be achievable if the algebraic Riccati equations for $X(\gamma)$ and $Y(\gamma)$ have solutions with $\rho(X(\gamma)Y(\gamma)) \leqq 1$. The optimal value of γ can occur when $\rho(X(\gamma)Y(\gamma)) = 1$ or when γ is the largest value such that the Hamiltonians for $X(\gamma)$ or $Y(\gamma)$ have imaginary axis eigenvalues. In this way the optimal value of γ can be calculated to any desired accuracy and a problem with $\rho(XY) = 1$ normally results.

4. Descriptor all-pass systems and the sufficient conditions. In this section we propose to lift the assumption that $\|R_a^{\sim}\|_H < 1$ and treat the case when $\|R_a^{\sim}\|_H = 1$. Before starting the general analysis, we consider a simple example which illustrates an important feature of the four block problem at optimality. Consider

$$\inf_{q \in \mathscr{H}_\infty^+} \left\| \begin{bmatrix} r_{11} & 0 \\ 0 & r_{22}+q \end{bmatrix} \right\|_\infty \tag{4.1}$$

in which $r_{11}, r_{22} \in \mathscr{H}_\infty^-$. It is immediate that there are two cases which require separate consideration. These are

$$(1) \quad \|r_{22}\|_H < \|r_{11}\|_\infty, \tag{4.2a}$$

$$(2) \quad \|r_{22}\|_H \geqq \|r_{11}\|_\infty. \tag{4.2b}$$

In the first case, any $q(s) \in \mathscr{H}_\infty^+$ for which $\|r_{22}+q\|_\infty \leqq \|r_{11}\|_\infty$ will be an optimal solution; a continuum of such solutions exist. In the second case, however, the (unique) optimal Nehari extension of $r_{22}(s)$ is the only solution. This example shows that there may or may not be a reduction in the size of the solution set at optimality. In the one block case, the size of the solution set always decreases at optimality [16].

In the general case there are also two forms of optimality. To see this we temporarily suppose that $Q(s)$ in (3.2) is allowed to range over $\mathscr{L}_\infty$. Under these conditions we have

$$\inf_{Q \in \mathscr{L}_\infty} \left\| \begin{matrix} R_{11} & R_{12} \\ R_{21} & R_{22}+Q \end{matrix} \right\|_\infty = \max\left\{ \left\| \begin{bmatrix} R_{11} \\ R_{21} \end{bmatrix} \right\|_\infty, \|[R_{11} \quad R_{12}]\|_\infty \right\} \tag{4.3}$$

by Parrott's theorem [39], [41, Thm. 1.2]; a particularly nice treatment of Parrott's theorem is given in Young [53]. The point is that in certain cases

$$\inf_{Q \in \mathscr{L}_\infty} \left\| \begin{matrix} R_{11} & R_{12} \\ R_{21} & R_{22}+Q \end{matrix} \right\|_\infty = \inf_{Q \in \mathscr{H}_\infty^+} \left\| \begin{matrix} R_{11} & R_{12} \\ R_{21} & R_{22}+Q \end{matrix} \right\|_\infty \tag{4.4}$$

and the requirement that Q be an element of $\mathscr{H}_\infty^+$ rather than just $\mathscr{L}_\infty$ makes no difference to the achievable norm. This Parrott type of optimality has already been covered by Theorem 3.6. The alternative form of optimality is treated below.

4.1. All optimal solutions: the general case. The purpose of this section is to treat sufficiency in the case that $\|R_a^{\sim}\|_H = 1$. We will take Lemma 3.4 as our starting point and then show that an all-pass embedding may still be constructed along the lines of Theorem 3.6. The key difficulty in the case of $\|R_a^{\sim}\|_H = 1$ is that Z in (3.65) is singular. Our approach to this problem will be to construct the all-pass embedding in a descriptor framework. We begin with an all-pass lemma for descriptor systems of the form (4.5) and (4.6) below. Apart from dealing with the standard case of a possibly singular E, we need to cater to the case of $\det(sE-A) \equiv 0$; $(sE-A)$ singular for all values of s.

THEOREM 4.1. *Consider the descriptor system of equations*

$$sEx(s) = Ax(s) + Bu(s), \tag{4.5}$$

$$y = Cx(s) + Du(s), \tag{4.6}$$

and suppose that there exists a matrix T such that

$$AT + T'A' + BB' = 0, \tag{4.7}$$

$$ET = T'E'. \tag{4.8}$$

Then

(a) *If T is nonsingular and*

$$CT = LB' \tag{4.9}$$

for some L, (4.5) *and* (4.6) *define a unique tranfer function given by*

$$G(s) = D + C(sE - A)^{\#}B \tag{4.10}$$

in which $(\cdot)^{\#}$ denotes a generalized inverse (which is defined in (4.16) *below).*

(b) *If T is nonsingular and*

$$\text{(i)}\quad DD' = I, \tag{4.11}$$

$$\text{(ii)}\quad CT + DB' = 0, \tag{4.12}$$

then

$$GG^{\sim} = I. \tag{4.13}$$

(c) *If $ET \geqq 0$ and $(sE - A, B)$ is stabilizable (i.e., $x'(sE - A) = 0, x'E \neq 0, x'B = 0 \Rightarrow (s + \bar{s}) < 0$), then all the finite eigenvalues of $sE - A$ satisfy $(s + \bar{s}) < 0$.*

Proof. In the proof of part (a), we use (4.7) and (4.8) to establish that $\mathcal{R}(B) \subset \mathcal{R}(sE - A)$. This allows (4.5) to be solved for $x(s)$ given any $u(s)$. Following that, (4.9) is used to prove that $\mathcal{N}(sE - A) \subset \mathcal{N}(C)$ which establishes the existence of the unique transfer function (4.10).

Suppose $sE - A$ has Smith diagonalization

$$sE - A = N(s)F(s)M(s) \tag{4.14}$$

in which $N(s)$ and $M(s)$ are unimodular polynomial matrices and

$$F(s) = \begin{bmatrix} F_1(s) & 0 \\ 0 & 0 \end{bmatrix}. \tag{4.15}$$

We define

$$(sE - A)^{\#} := M^{-1}(s)\begin{bmatrix} F_1^{-1}(s) & 0 \\ 0 & 0 \end{bmatrix} N^{-1}(s). \tag{4.16}$$

To show that (4.5) has a solution for $x(s)$ given any $u(s)$, we note that (4.7) and (4.8) give

$$(sE - A)T + T'(-sE' - A') = BB' \tag{4.17}$$

and hence from (4.14)

$$\begin{aligned} &FMT(N^{\sim})^{-1} + N^{-1}T'M^{\sim}F^{\sim} = N^{-1}BB'(N^{\sim})^{-1} \\ &\Rightarrow [0 \quad I]N^{-1}B = 0. \end{aligned} \tag{4.18}$$

It is clear from (4.18) that $\mathcal{R}(B) \subset \mathcal{R}(sE - A)$ and thus that (4.5) has a solution for all $u(s)$. A simple verification shows that $x(s)$ solves (4.5) if and only if

$$\begin{aligned} x(s) &= M^{-1}(s)\left\{\begin{bmatrix} F_1^{-1} & 0 \\ 0 & 0 \end{bmatrix} N^{-1}(s)Bu(s) + \begin{bmatrix} 0 \\ I \end{bmatrix} w(s)\right\} \\ &= (sE - A)^{\#}Bu(s) + M^{-1}(s)\begin{bmatrix} 0 \\ I \end{bmatrix} w(s) \end{aligned} \tag{4.19}$$

for some $w(s)$. Next, we note that (4.17) gives

$$\begin{aligned}
&(T')^{-1}(sE-A)+(-sE'-A)T^{-1}=(T')^{-1}BB'T^{-1}\\
&\Rightarrow (M^{\sim})^{-1}(T')^{-1}NF+F^{\sim}N^{\sim}T^{-1}M^{-1}=(M^{\sim})^{-1}(T')^{-1}BB'T^{-1}M^{-1}\\
&\Rightarrow B'T^{-1}M^{-1}\begin{bmatrix}0\\ I\end{bmatrix}=0.
\end{aligned}\tag{4.20}$$

Thus

$$\begin{aligned}
y(s)&=Cx(s)+Du(s)\\
&=Du(s)+C(sE-A)^{\#}Bu(s)+CM^{-1}\begin{bmatrix}0\\ I\end{bmatrix}w(s)\\
&=\{D+C(sE-A)^{\#}B\}u(s)+LB'T^{-1}M^{-1}\begin{bmatrix}0\\ I\end{bmatrix}w(s)\\
&=G(s)u(s)
\end{aligned}$$

by (4.9) and (4.20).

(b) Equations (4.12), (4.14), and (4.16) imply that

$$\begin{aligned}
C(sE-A)^{\#}(sE-A)&=-DB'T^{-1}M^{-1}\begin{bmatrix}I & 0\\ 0 & 0\end{bmatrix}M\\
&=-DB'T^{-1}\quad\text{by (4.20)}\\
&=C\quad\text{by (4.12).}
\end{aligned}\tag{4.21}$$

Thus

$$\begin{aligned}
I-GG^{\sim}=I-DD'-C(sE-A)^{\#}BD'-DB'(-sE-A')^{\#}C'&\\
-C(sE-A)^{\#}((sE-A)T+T'(-sE'-A'))&\\
\times(-sE-A')^{\#}C'\quad\text{by (4.10) and (4.17)}&\\
=0\quad\text{by (4.11), (4.12), and (4.21).}&
\end{aligned}\tag{4.22}$$

(c) Let λ be a finite eigenvalue of $(sE-A)$. Then there exists a corresponding eigenvector x such that (i) $x'E\neq 0$ and $x'(\lambda E-A)=0$. $x'(4.7)x\Rightarrow(\lambda+\bar{\lambda})x'ETx=-x'BB'x\leqq 0$. If $x'B\neq 0$, then $\lambda+\bar{\lambda}<0$ since $ET\geqq 0$. If $x'B=0$, then $\lambda+\bar{\lambda}<0$ by the stabilizability assumption. □

Proceeding as before, we postulate matrices T and E which are associated with equations (4.7) and (4.8), and which satisfy (4.12)

$$T=\begin{bmatrix}-X & Z\\ I & Y\end{bmatrix},\qquad E=\begin{bmatrix}I & 0\\ 0 & Z'\end{bmatrix}\tag{4.23}$$

where X, Y, and Z are defined in (3.51), (3.52), and (3.65). Equations (3.67) and (3.68) become

$$\begin{bmatrix}A & 0\\ 0 & \hat{A}_0\end{bmatrix}\begin{bmatrix}-X & Z\\ I & Y\end{bmatrix}+\begin{bmatrix}-X & I\\ Z' & Y\end{bmatrix}\begin{bmatrix}A' & 0\\ 0 & \hat{A}_0'\end{bmatrix}+\begin{bmatrix}B_e\\ \hat{B}_{e0}\end{bmatrix}[B_e'\quad \hat{B}_{e0}']=0\tag{4.24}$$

and

$$D_e[B_e'\quad \hat{B}_{e0}']+[C_e\quad \hat{C}_{e0}]\begin{bmatrix}-X & Z\\ I & Y\end{bmatrix}=0,\tag{4.25}$$

respectively. Evaluating the first column of (4.25) and comparing with (3.69) gives

$$\hat{C}_{e0} = \hat{C}_e = C_eX - D_eB'_e. \tag{4.26}$$

In the same way, the second column of (4.25) together with (3.70) yields

$$\hat{B}_{e0} = YB_e - C'_eD_e = Z'\hat{B}_e \tag{4.27}$$

for any unitary D_e. $\hat{B}_{e0}$ therefore retains the critical zero (1, 1) entry. The (1, 1) partition of (4.24) is zero by (3.51). From the (2, 1) partition of (4.24) we obtain

$$\hat{A}_0 = -Z'A' - \hat{B}_{e0}B'_e = -A' - YAX + C'_eD_eB'_e = Z'\hat{A}. \tag{4.28}$$

It is now easy to check that the remaining equations in (4.24) are also satisfied.

We also note that $(Z', \hat{A}_0, \hat{B}_{e0})$ is stabilizable since if $x'Z' \neq 0$, $x'[sZ' - \hat{A}_0, \hat{B}_{e0}] = 0 \Rightarrow x'[sZ' + Z'A'] = 0 \Rightarrow s + \bar{s} < 0$ since Re $(\lambda_i(A)) > 0$. We can now prove the sufficiency of the condition in Corollary 3.5(i).

THEOREM 4.2. *Given that the conditions of Corollary* 3.5(i) *are satisfied, then*

(a) $Q_{aa}(s) := D_e + \hat{C}_{e0}(sZ' - \hat{A}_0)^{\#}\hat{B}_{e0} \in \mathcal{RH}_\infty$,
(b) $R_{aa} + Q_{aa}$ *is all-pass.*
Hence the conditions are both necessary and sufficient.

Proof. Equations (4.24) and (4.25) give that Theorem 4.1 can be applied to show that $R_{aa} + Q_{aa}$ is well defined and all-pass. Since $R_{aa}^{\sim} \in \mathcal{RH}_\infty$, Q_{aa} has no poles on $s = j\omega$ or at infinity. All the finite poles will be at eigenvalues of $(sZ' - \hat{A}_0)$ which are in the open left halfplane by Theorem 4.1(c) since $(Z', \hat{A}_0, \hat{B}_{e0})$ is stabilizable and from the (2, 2)-block of (4.24) and $Z'Y = Y - YXY \geqq 0$ since $\rho(XY) \leqq 1$. The sufficiency of the condition follows from (a) and (b) by using the (2, 2) entry of $Q_{aa}(s)$. □

4.2. Characterization of all solutions. Theorem 4.2 gives a solution in the case $\|R_a\|_H = 1$ in descriptor form. All solutions can be generated from $Q_{aa}(s)$ but the more detailed structure is required. To keep the notation simple the simplifying assumption

$$\begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

will be made for this section. Model matching problems arising from H^∞ control can in fact always be reduced to this case (see [45]). The main difference in the approach in § 3 is that all solutions to the problem with R_a rather than R are determined. We wish to write any solution

$$\begin{bmatrix} \bar{Q}_{22} & \bar{Q}_{23} \\ \bar{Q}_{32} & \bar{Q}_{33} \end{bmatrix}$$

as

$$\begin{bmatrix} \bar{Q}_{22} & \bar{Q}_{23} \\ \bar{Q}_{32} & \bar{Q}_{33} \end{bmatrix} = \begin{bmatrix} Q_{22} & Q_{23} \\ Q_{32} & Q_{33} \end{bmatrix} + \begin{bmatrix} \tilde{Q}_{24} \\ \tilde{Q}_{34} \end{bmatrix} \Phi(I - \tilde{Q}_{44}\Phi)^{-1}[\tilde{Q}_{42} \quad \tilde{Q}_{43}]$$

for $\|\Phi\|_\infty \leqq 1$, $\Phi \in H_\infty^+$. We need to show that such a Φ exists for all suitable $\bar{Q}_{ij}$, and this is not immediately clear since

$$\begin{bmatrix} \tilde{Q}_{24} \\ \tilde{Q}_{34} \end{bmatrix} \quad \text{and} \quad [\tilde{Q}_{42} \quad \tilde{Q}_{43}]$$

are not invertible. The following technical lemma will be used to check the existence of a Φ given any

$$\begin{bmatrix} \bar{Q}_{22} & \bar{Q}_{23} \\ \bar{Q}_{32} & \bar{Q}_{33} \end{bmatrix}.$$

LEMMA 4.3. *Suppose that*

$$P(s) = \begin{array}{c c} & \begin{array}{cc} m_1 & m_2 \end{array} \\ \begin{array}{c} p_1 \\ p_2 \end{array} & \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \end{array} (s) \in R\mathcal{L}_\infty,\ X \in R\mathcal{L}_\infty^{p_1 \times m_1}$$

in which $p_1 \geqq m_2$, $m_1 \geqq p_2$ *and* $P_{22}(\infty) = 0$. *Suppose also that* P_{12} *has a left inverse* $P_{12}^l \in \mathcal{L}_\infty$ *and* P_{21} *has a right inverse* $P_{21}^r \in \mathcal{L}_\infty$. *Then if there exists a rational matrix* $R \in \mathcal{L}_\infty$ *with* rank $(R) \geqq p_1 - m_2$ *for almost all s such that*

$$\text{(4.29)} \qquad \text{(i)} \quad R[X - P_{11} \quad P_{12}] = 0,$$

and if there exists a rational matrix $S \in \mathcal{L}_\infty$ *with* rank $(S) \geqq m_1 - p_2$ *for almost all s such that*

$$\text{(4.30)} \qquad \text{(ii)} \quad \begin{bmatrix} X - P_{11} \\ P_{21} \end{bmatrix} S = 0,$$

then there exists a rational matrix Φ *such that* $F_l(P, \Phi) = X$. *More particularly,*

$$\text{(4.31)} \qquad \Phi = (I + \Psi P_{22})^{-1} \Psi$$

where

$$\text{(4.32)} \qquad \Psi = P_{12}^l (X - P_{11}) P_{21}^r.$$

Proof. The idea of the proof is to establish that (i) and (ii) guarantee the existence of a solution Ψ to the equation $X - P_{11} = P_{12} \Psi P_{21}$. We then set $\Psi = \Phi(I - P_{22}\Phi)^{-1}$ and solve for Φ.

It follows from (i) that $(X - P_{11}) \in \mathcal{N}(R)$. Since $\dim \mathcal{R}(P_{12}) = m_2 \geqq \dim \mathcal{N}(R)$, it follows from (i) that $(X - P_{11}) \in \mathcal{N}(R) \equiv \mathcal{R}(P_{12})$ and therefore that the equation $P_{12} Z = (X - P_{11})$ has a solution $Z = P_{12}^l (X - P_{11})$. In the same way we have $(X - P_{11})^\sim \in \mathcal{R}(P_{21}^\sim)$. Since $Z^\sim \in \mathcal{R}(X - P_{11})^\sim$, we have $Z^\sim \in \mathcal{R}(P_{21}^\sim)$ and consequently $\Psi P_{21} = Z$ has a solution $\Psi = P_{12}^l (X - P_{11}) P_{21}^r \in \mathcal{L}_\infty$. It follows from $P_{22}(\infty) = 0$ and $\Psi = \Phi(I - P_{22}\Phi)^{-1}$ that $\Phi = (I + \Psi P_{22})^{-1} \Psi$ is a proper rational matrix. □

The detailed structure of $Q_{aa}(s)$ will now be examined when

$$\begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} = 0.$$

In this case we may choose

$$\text{(4.33)} \qquad D_e = \begin{bmatrix} 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \\ I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \end{bmatrix}$$

and substituting (4.33) into (4.26)-(4.28) yields

$$\text{(4.34)} \qquad \hat{C}'_{e0} = [0 \quad XC'_2 \quad XC'_3 - B_1 \quad -B_2],$$

$$\hat{B}_{e0} = [0 \quad YB_2 \quad YB_3 - C_1' \quad -C_2'], \tag{4.35}$$

$$\hat{A}_0 = -A' - YAX + C_3'B_1' + C_1'B_3'. \tag{4.36}$$

Since (3.51) and (3.52) have an appropriate dual relationship to each other, their solutions may be transformed to the balanced form

$$X = \begin{bmatrix} \bar{X} & 0 \\ 0 & I_k \end{bmatrix}, \quad Y = \begin{bmatrix} \bar{Y} & 0 \\ 0 & I_k \end{bmatrix}, \quad \bar{Z} = I - \bar{X}\bar{Y} \tag{4.37}$$

by an appropriate change of basis in the state space of $R(s)$; k is the multiplicity of the unit singular value of $\Gamma_{R_a^\sim}$. This balancing induces the following partitioning on B_e and C_e in (3.59) and (3.60).

$$\begin{bmatrix} B_{e1} \\ B_{e2} \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} & B_{13} & 0 \\ B_{21} & B_{22} & B_{23} & 0 \end{bmatrix}, \tag{4.38}$$

$$\begin{bmatrix} C_{e1}' \\ C_{e2}' \end{bmatrix} = \begin{bmatrix} C_{11}' & C_{12}' & C_{13}' & 0 \\ C_{21}' & C_{22}' & C_{23}' & 0 \end{bmatrix}. \tag{4.39}$$

Furthermore if rank $B_{22} = l < k$, then an additional unitary change of basis on the k states can give $B_{22} = \left[\begin{smallmatrix} B_{221} \\ 0 \end{smallmatrix}\right]$, with B_{221} full rank.

Substitution into (3.49) and (3.50) gives

$$\begin{bmatrix} B_{13} \\ B_{23} \end{bmatrix} = \begin{bmatrix} \bar{X}C_{11}' \\ C_{21}' \end{bmatrix}, \tag{4.40}$$

$$\begin{bmatrix} C_{13}' \\ C_{23}' \end{bmatrix} = \begin{bmatrix} \bar{Y}B_{11} \\ B_{21} \end{bmatrix}. \tag{4.41}$$

Finally, (3.51) and (3.52) become

(4.42)

$$-\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}\begin{bmatrix} \bar{X} & 0 \\ 0 & I_k \end{bmatrix} - \begin{bmatrix} \bar{X} & 0 \\ 0 & I_k \end{bmatrix}\begin{bmatrix} A_{11}' & A_{21}' \\ A_{12}' & A_{22}' \end{bmatrix} + \begin{bmatrix} B_{11} & B_{12} & \bar{X}C_{11}' \\ B_{21} & B_{22} & C_{21}' \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} & \bar{X}C_{11}' \\ B_{21} & B_{22} & C_{21}' \end{bmatrix}' = 0,$$

(4.43)

$$-\begin{bmatrix} A_{11}' & A_{21}' \\ A_{12}' & A_{22}' \end{bmatrix}\begin{bmatrix} \bar{Y} & 0 \\ 0 & I_k \end{bmatrix} - \begin{bmatrix} \bar{Y} & 0 \\ 0 & I_k \end{bmatrix}\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} + \begin{bmatrix} C_{11}' & C_{12}' & \bar{Y}B_{11} \\ C_{21}' & C_{22}' & B_{21} \end{bmatrix} \begin{bmatrix} C_{11}' & C_{12}' & \bar{Y}B_{11} \\ C_{21}' & C_{22}' & B_{21} \end{bmatrix}' = 0.$$

The (2, 2) blocks of (4.42) and (4.43) establish the following equation:

$$B_{22}B_{22}' = C_{22}'C_{22} \tag{4.44}$$

and consequently that there exists a matrix U such that

$$UB_{22}' = C_{22}, \qquad U := C_{22}(B_{22}')^{\dagger} = (C_{22}')^{\dagger}B_{22}. \tag{4.45}$$

It now follows that a descriptor representation for $Q_{aa}(s)$ is given by

$$\begin{bmatrix} \bar{Z}' & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} \dot{x}_1 \\ \dot{x}_2 \end{bmatrix} = \begin{bmatrix} -A'_{11} - \bar{Y}A_{11}\bar{X} + C'_{13}B'_{11} + C'_{11}B'_{13} & -C'_{12}UB'_{22} \\ -B_{22}B'_{12} & -B_{22}B'_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 & \bar{Y}B_{12} & -\bar{Z}'C'_{11} & -C'_{12} \\ 0 & B_{22} & 0 & -B_{22}U' \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix}, \tag{4.46}$$

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ C_{12}\bar{X} & UB'_{22} \\ -B'_{11}\bar{Z}' & 0 \\ -B'_{12} & -B'_{22} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \\ I & 0 & 0 & 0 \\ 0 & I & 0 & 0 \end{bmatrix}\begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix} \tag{4.47}$$

on substituting (4.37) to (4.41) into (4.34)-(4.36). The (2, 1) block of (4.46) gives

$$\begin{aligned} 0 &= -B_{22}B'_{12}x_1 - B_{22}B'_{22}x_2 + B_{22}u_2 - B_{22}U'u_4 \\ &\Rightarrow B'_{22}x_2 = (B_{22})^{\dagger}(-B_{22}B'_{12}x_1 + B_{22}u_2 - C'_{22}u_4). \end{aligned} \tag{4.48}$$

Substituting (4.48) into (4.46) and (4.47) gives the state-space realization,

$$\begin{bmatrix} Q_{22} & Q_{23} & Q_{24} \\ Q_{32} & Q_{33} & Q_{34} \\ Q_{42} & Q_{43} & Q_{44} \end{bmatrix}(s) \overset{s}{=} \left[\begin{array}{c|ccc} \tilde{A} & (\bar{Z}')^{-1}(\bar{Y}B_{12} - C'_{12}U) & -C'_{11} & -(\bar{Z}')^{-1}C'_{12}(I - C_{22}C^{\dagger}_{22}) \\ \hline C_{12}\bar{X} - UB'_{12} & U & 0 & (I - C_{22}C^{\dagger}_{22}) \\ -B'_{11}\bar{Z}' & 0 & 0 & 0 \\ -(I - B^{\dagger}_{22}B_{22})B'_{12} & I - B^{\dagger}_{22}B_{22} & 0 & -B^{\dagger}_{22}C'_{22} \end{array}\right] \tag{4.49}$$

where

$$\tilde{A} = (\bar{Z}')^{-1}(-A'_{11} - \bar{Y}A_{11}\bar{X} + C'_{13}B'_{11} + C'_{11}B'_{13} + C'_{12}C_{22}(B'_{22})^{\dagger}B'_{12}).$$

Suppose that Θ_1 and Θ_3 are orthogonal bases for $\mathscr{R}(B'_{22})$ and $\mathscr{R}(C_{22})$, respectively, and that Θ_2 and Θ_4 are chosen to make $[\Theta_2 \quad \Theta_1]$ and $[\Theta_4 \quad \Theta_3]$ orthogonal.[1] If we multiply the last row of (4.49) by $[\begin{smallmatrix}\Theta'_2 \\ \Theta'_1\end{smallmatrix}]$ and the last column by $[\Theta_4 \quad \Theta_3]$, we obtain

$$\begin{bmatrix} Q_{22} & Q_{23} & \tilde{Q}_{24} & 0 \\ Q_{32} & Q_{33} & \tilde{Q}_{34} & 0 \\ \tilde{Q}_{42} & \tilde{Q}_{43} & \tilde{Q}_{44} & 0 \\ 0 & 0 & 0 & -I_l \end{bmatrix}(s) \tag{4.50}$$

[1] If B'_{22} has a singular value decomposition

$$B'_{22} = [Y_1 \quad Y_2]\begin{bmatrix} 0 & 0 \\ 0 & \Gamma \end{bmatrix}\begin{bmatrix} U'_1 \\ U'_2 \end{bmatrix},$$

then $[\Theta_2 \quad \Theta_1] = [Y_1 \quad Y_2]$ and $(B'_{22})^{\dagger} = U_2\Gamma^{-1}Y'_2$.

$$\overset{s}{=}\left[\begin{array}{c|cccc} \tilde{A} & (\bar{Z}')^{-1}(\bar{Y}B_{12}-C_{12}'U) & -C_{11}' & -(\bar{Z}')^{-1}C_{12}'\Theta_4 & 0 \\ \hline C_{12}\bar{X}-UB_{12}' & U & 0 & \Theta_4 & 0 \\ -B_{11}'\bar{Z}' & 0 & 0 & 0 & 0 \\ -\Theta_2'B_{12}' & \Theta_2' & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -I_l \end{array}\right]$$

in which

$$l = \operatorname{rank}(B_{22}) = \operatorname{rank}(C_{22}) \leqq k. \tag{4.51}$$

Equation (4.50) has an interesting structure in that it shows how the effective dimension of the free parameter is reduced by l. We will examine this point a little further and obtain a constraining equation on the free parameter Φ which is reminicent of [16, eq. (8.69)]. Following that, we will prove that

(4.52)

$$\begin{bmatrix} Q_{22} & \tilde{Q}_{24} \\ \tilde{Q}_{42} & \tilde{Q}_{44} \end{bmatrix}(s) \overset{s}{=} \left[\begin{array}{c|cc} \tilde{A} & (\bar{Z}')^{-1}(\bar{Y}B_{12}-C_{12}'U) & -(\bar{Z}')^{-1}C_{12}'\Theta_4 \\ \hline C_{12}\bar{X}-UB_{12}' & U & \Theta_4 \\ -\Theta_2'B_{12}' & \Theta_2' & 0 \end{array}\right]$$

captures all solutions in the optimal case corresponding to $\|\Gamma_{R_a^\sim}\| = 1$.

Let us write the left-hand side of (4.50) as

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} Q_{22} & \tilde{Q}_{24} & 0 \\ \tilde{Q}_{42} & \tilde{Q}_{44} & 0 \\ 0 & 0 & I_l \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \tag{4.53}$$

and allow this to induce the following partitioning on the free contraction $\Phi(s)$

$$\begin{bmatrix} u_2 \\ u_3 \end{bmatrix} = \begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix} \begin{bmatrix} y_2 \\ y_3 \end{bmatrix}. \tag{4.54}$$

Since $u_3 = y_3$ we may eliminate this variable in (4.54) to obtain

$$\begin{aligned} u_2 &= (\Phi_{11} + \Phi_{12}(I-\Phi_{22})^{-1}\Phi_{21})y_2 \quad \text{if } \det(I-\Phi_{22}(\infty)) \neq 0 \text{ is assumed} \\ &= \left(F_l\left\{\begin{bmatrix} \Phi_{11} & \Phi_{12} \\ \Phi_{21} & \Phi_{22} \end{bmatrix}, I_l\right\}\right) y_2 = \tilde{\Phi} y_2 \end{aligned} \tag{4.55}$$

in which $\|\tilde{\Phi}\| \leqq 1$ by Redheffer's theorem [42]. A class of solutions is therefore given by

$$Q = F_l\left\{\begin{bmatrix} Q_{22} & \tilde{Q}_{24}' \\ \tilde{Q}_{42} & \tilde{Q}_{44} \end{bmatrix}, \tilde{\Phi}\right\}. \tag{4.56}$$

Transforming back to the coordinates in (4.53) gives

$$\Phi = U[\Theta_2 \quad \Theta_1] \begin{bmatrix} \tilde{\Phi} & 0 \\ 0 & I_l \end{bmatrix} \begin{bmatrix} \Theta_2' \\ \Theta_1 \end{bmatrix} \tag{4.57}$$

and $(4.57)B_{22}'$

$$\Rightarrow \Phi B_{22}' = C_{22}, \tag{4.58}$$

which is a linear constraint on Φ and similar to that obtained in [16].

We are now in a position to state and prove the main result of this section.

THEOREM 4.4. *Let $R_{ij} \in \mathcal{RH}_\infty^{-,p_i,m_j}$ for $i, j = 1, 2$ have the realization*

$$\begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \overset{s}{=} \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0 & 0 \\ C_2 & 0 & 0 \end{array}\right]$$

with $\operatorname{Re} \lambda_i(A) > 0$.

If the conditions of Corollary 3.5(i) *are met, then every* $Q \in \mathcal{RH}_\infty^+$ *satisfying*

$$\left\| \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} + Q \end{bmatrix} \right\|_\infty \leqq 1$$

is given by

$$Q = F_l\left(\begin{bmatrix} Q_{22} & \tilde{Q}_{24} \\ \tilde{Q}_{42} & \tilde{Q}_{44} \end{bmatrix}, \Phi \right) \quad \textit{for some } \Phi \in \mathcal{RH}_\infty^+ \textit{ with } \|\Phi\|_\infty \leqq 1 \tag{4.59}$$

in which

$$\begin{bmatrix} Q_{22} & \tilde{Q}_{24} \\ \tilde{Q}_{42} & \tilde{Q}_{44} \end{bmatrix}$$

has a state-space realization given by (4.52). □

Proof. The proof can be deduced using arguments similar to those given in the proof of Theorem 3.6. We begin by making two preliminary observations. First, it is an immediate consequence of Redheffers theorem that (4.59) captures a class of solutions. Second,

$$\begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} + \bar{Q}_{22} & R_{23} + \bar{Q}_{23} \\ R_{31} & R_{32} + \bar{Q}_{32} & R_{33} + \bar{Q}_{33} \end{bmatrix} \begin{bmatrix} R_{11}^\sim & V_{11} \\ R_{12}^\sim & V_{21} \\ R_{13}^\sim & V_{31} \end{bmatrix} = \begin{bmatrix} I & U_{11}(-s) \\ 0 & U_{21}(-s) \\ 0 & U_{31}(-s) \end{bmatrix} \tag{4.60}$$

for any

$$Q = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \bar{Q}_{22} & \bar{Q}_{23} \\ 0 & \bar{Q}_{32} & \bar{Q}_{33} \end{bmatrix}$$

such that $\|R_a + Q\|_\infty = \|R_a^\sim\|_H = 1$ where $V(s)$ and $U(-s)$ are the maximal Schmidt vectors of the Hankel operator $\Gamma_{R_a^\sim}$ [14], [18]. Since

$$\begin{bmatrix} R_{11} & R_{12} & R_{13} \\ V_{11}^\sim & V_{21}^\sim & V_{31}^\sim \end{bmatrix}(\infty) = \begin{bmatrix} 0 & 0 & I \\ C_{21} & C_{22} & C_{23} \end{bmatrix},$$

it is immediate that

$$p_1 + k \geqq \operatorname{rank} \begin{bmatrix} R_{12}^\sim & V_{21} \\ R_{13}^\sim & V_{31} \end{bmatrix} \geqq p_1 + \operatorname{rank}(C_{22}) \tag{4.61}$$

for almost all s; k is the multiplicity of the unit Hankel singular value of $\Gamma_{R_a^\sim}$. Similarly, since $\|R_{aa} + Q_{aa}\|_\infty = \|R_{aa}^\sim\|_H = 1$

$$\begin{bmatrix} 0 & 0 & 0 \\ 0 & \bar{Q}_{22} - Q_{22} & \bar{Q}_{23} - Q_{23} \\ 0 & \bar{Q}_{32} - Q_{32} & \bar{Q}_{33} - Q_{33} \\ 0 & \tilde{Q}_{42} & \tilde{Q}_{43} \end{bmatrix} \begin{bmatrix} R_{11}^\sim & V_{11} \\ R_{12}^\sim & V_{21} \\ R_{13}^\sim & V_{31} \end{bmatrix} = 0. \tag{4.62}$$

In the same way we have that

$$(4.63)\qquad \begin{bmatrix} R_{11}^{\sim} & R_{21}^{\sim} & R_{31}^{\sim} \\ U_{11}^{\sim} & U_{21}^{\sim} & U_{31}^{\sim} \end{bmatrix} \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & \bar{Q}_{22}-Q_{22} & \bar{Q}_{23}-Q_{23} & \tilde{Q}_{24} \\ 0 & \bar{Q}_{32}-Q_{32} & \bar{Q}_{33}-Q_{33} & \tilde{Q}_{34} \end{bmatrix} = 0$$

and

$$(4.64)\qquad \operatorname{rank} \begin{bmatrix} R_{11}^{\sim} & R_{21}^{\sim} & R_{31}^{\sim} \\ U_{11}^{\sim} & U_{21}^{\sim} & U_{31}^{\sim} \end{bmatrix} \geqq m_1 + \operatorname{rank}(B_{22}).$$

We now show that $[\begin{smallmatrix}\tilde{Q}_{24}\\ \tilde{Q}_{34}\end{smallmatrix}]$ has an asymptotically stable left inverse while $[\tilde{Q}_{42} \quad \tilde{Q}_{43}]$ has an asymptotically stable right inverse. Substituting from (4.50) gives

$$(4.65)\qquad \begin{bmatrix} \tilde{Q}_{24} \\ \tilde{Q}_{34} \end{bmatrix} \overset{s}{=} \left[\begin{array}{c|c} \tilde{A} & -(\bar{Z}')^{-1}C_{12}'\Theta_4 \\ \hline C_{12}\bar{X}-UB_{12}' & \Theta_4 \\ -B_{11}'\bar{Z}' & 0 \end{array}\right]$$

and a simple calculation yields[2]

$$(4.66)\qquad \begin{bmatrix} \tilde{Q}_{24} \\ \tilde{Q}_{34} \end{bmatrix}^{l} \overset{s}{=} \left[\begin{array}{c|c} -(\bar{Z}')^{-1}A_{11}\bar{Z} & (\bar{Z}')^{-1}[C_{12}' \quad C_{13}'] \\ \hline \Theta_4'[C_{12}\bar{X}-UB_{12}'] & \Theta_4' \quad 0 \end{array}\right].$$

$$(4.67)\qquad [\tilde{Q}_{42} \quad \tilde{Q}_{43}] \overset{s}{=} \left[\begin{array}{c|cc} \tilde{A} & (\bar{Z}')^{-1}[\bar{Y}B_{12}-C_{12}'U]-C_{11}' & \\ \hline -[\Theta_2' \quad 0]\begin{bmatrix} B_{12}' \\ B_{13}' \end{bmatrix} & \Theta_2' & 0 \end{array}\right]$$

has a right inverse

$$(4.68)\qquad [\tilde{Q}_{42} \quad \tilde{Q}_{43}]^{r} \overset{s}{=} \left[\begin{array}{c|c} -A_{11}' & -(\bar{Z}')^{-1}[\bar{Y}B_{12}-C_{12}'U]\Theta_2 \\ \hline B_{12}' & \Theta_2 \\ B_{13}' & 0 \end{array}\right].$$

Since $\lambda_i(A_{11})>0$, the left and right inverses in (4.66) and (4.68) are asymptotically stable. Furthermore,

$$(4.69)\qquad \operatorname{rank} \begin{bmatrix} \tilde{Q}_{24} \\ \tilde{Q}_{34} \end{bmatrix} = p_2 - \operatorname{rank}(B_{22}) \quad \text{for all } \operatorname{Re}(s) \geqq 0,$$

$$(4.70)\qquad \operatorname{rank}[\tilde{Q}_{42} \quad \tilde{Q}_{43}] = m_2 - \operatorname{rank}(B_{22}) \quad \text{for all } \operatorname{Re}(s) \geqq 0.$$

Hence, we may invoke Lemma 4.3 and the all-pass character of E_{aa} to establish the existence of a $\|\Phi\|_\infty \leqq 1$ such that

$$(4.71)\qquad \begin{bmatrix} \bar{Q}_{22} & \bar{Q}_{23} \\ \bar{Q}_{32} & \bar{Q}_{33} \end{bmatrix} = F_l\left(\begin{bmatrix} Q_{22} & Q_{23} & \tilde{Q}_{24} \\ Q_{32} & Q_{33} & \tilde{Q}_{34} \\ \tilde{Q}_{42} & \tilde{Q}_{43} & \tilde{Q}_{44} \end{bmatrix} ; \Phi \right).$$

It remains for us to show that any such $\Phi \in \mathscr{BH}_\infty^+$. It follows immediately that

$$(4.72)\qquad \Psi = \begin{bmatrix} \tilde{Q}_{24} \\ \tilde{Q}_{34} \end{bmatrix}^{l} \begin{bmatrix} \bar{Q}_{22}-Q_{22} & \bar{Q}_{23}-Q_{23} \\ \bar{Q}_{32}-Q_{32} & \bar{Q}_{33}-Q_{33} \end{bmatrix} [\tilde{Q}_{42} \quad \tilde{Q}_{43}]^{r}$$

[2] $$G(s) \overset{s}{=} \begin{bmatrix} A & BD \\ C & D \end{bmatrix} \Rightarrow G^{l}(s) \overset{s}{=} \begin{bmatrix} A-BC & -B \\ D^{l}C & D^{l} \end{bmatrix}$$

and a dual result is clearly true in the case of a right inverse.

is stable. Now suppose that Φ has a coprime factorization $\Phi = ND^{-1}$ with $N, D \in \mathcal{RH}_\infty^+$. Since $\Psi = \Phi(I - Q_{44}\Phi)^{-1}$ is stable, and $\Psi = N(D - Q_{44}N)^{-1}$ is a coprime factorization, it follows that $(D - Q_{44}N)$ is outer. We may now deduce from

$$\det(D - \tilde{Q}_{44}N)(j\omega) = \det D(j\omega) \cdot \det(I - \tilde{Q}_{44}\Phi)(j\omega) \tag{4.73}$$

and $\|\tilde{Q}_{44}\Phi\|_\infty < 1$ that the winding number of $\det D(j\omega)$ around the origin is zero. This means that D is outer and therefore that $\Phi = ND^{-1} \in \mathcal{RH}_\infty^+$ as required. □

5. A representation formula for all internally stabilizing controllers that satisfy a closed-loop H^∞ norm constraint. In this section we will show how the solution to the four block problem given in §§ 3 and 4 may be used to solve a rational $\mathcal{H}^\infty$ control problem. Suppose we are given

$$P = \begin{array}{cc} \begin{array}{cc} l & m \end{array} & \\ \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} & \begin{array}{c} p \\ q \end{array} \end{array} \overset{s}{=} \begin{array}{cc} \begin{array}{ccc} n & l & m \end{array} & \\ \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & D_{11} & D_{12} \\ C_2 & D_{21} & D_{22} \end{array}\right] & \begin{array}{c} n \\ p \\ q \end{array} \end{array} \tag{5.1}$$

in which $p \geqq m$ and $l \geqq q$, and that we seek all controllers that stabilize $F_l(P, K)$ and that satisfy the norm constraint

$$\|F_l(P, K)\|_\infty \leqq \gamma. \tag{5.2}$$

We will make a number of assumptions regarding (5.1), the last of which is temporary and is removed in our later development.

(i) (A, B_2, C_2) is stabilizable and detectable.

(ii) $\operatorname{rank}(D_{12}) = m$ and $\operatorname{rank}(D_{21}) = q$.

(iii) $\operatorname{rank}\left(\begin{bmatrix} j\omega I - A & -B_2 \\ C_1 & D_{12} \end{bmatrix}\right) = n + m$ for all real ω.

(iv) $\operatorname{rank}\left(\begin{bmatrix} j\omega I - A & -B_1 \\ C_2 & D_{21} \end{bmatrix}\right) = n + q$ for all real ω.

(v) *Temporary assumption.* $D_{11} = 0$ and $D_{22} = 0$.

Additionally, the results that follow presume that the problem has been scaled so that the columns of D_{12} and D'_{21} are orthogonal. This is always possible by assumption (ii) (see [43]). We will introduce $D_\perp$ and $\tilde{D}_\perp$ which make $[D_\perp D_{12}]$ and $[\tilde{D}'_\perp D'_{21}]$ unitary.

The main results of this section will now be stated, and their proofs will be given in the following three sections. Theorem 5.1 gives necessary and sufficient conditions for the existence of a solution, while Theorem 5.2 characterizes all the solutions. Assumption (v) will be removed in § 5.3.

THEOREM 5.1. *Suppose that $P(s)$ is given by* (5.1) *and that assumptions* (i)–(v) *are satisfied. Then for any $\gamma > 0$ there exists an internally stabilizing controller $K(s)$ such that $\|F_l(P, K)\|_\infty \leqq \gamma$, if and only if*:

(i) *There exists*

$$\begin{bmatrix} X_{\infty 1} \\ X_{\infty 2} \end{bmatrix} \in \mathcal{R}^{2n \times n} \text{ of } \operatorname{rank}(n)$$

such that

$$H_\infty \begin{bmatrix} X_{\infty 1} \\ X_{\infty 2} \end{bmatrix} = \begin{bmatrix} X_{\infty 1} \\ X_{\infty 2} \end{bmatrix} T_x, \quad \operatorname{Re} \lambda_i(T_x) \leqq 0 \quad \forall i, \tag{5.3}$$

$$X'_{\infty 1} X_{\infty 2} = X'_{\infty 2} X_{\infty 1} \tag{5.4}$$

where

(5.5) $$H_\infty = \begin{bmatrix} A - B_2 D'_{12} C_1 & \gamma^{-2} B_1 B'_1 - B_2 B'_2 \\ -C'_1 D_\perp D'_\perp C_1 & -(A - B_2 D'_{12} C_1)' \end{bmatrix}.$$

(ii) *There exists*

$\begin{bmatrix} Y_{\infty 1} \\ Y_{\infty 2} \end{bmatrix} \in \mathcal{R}^{2n \times n}$ *of* rank (n) *such that*

(5.6) $$J_\infty \begin{bmatrix} Y_{\infty 1} \\ Y_{\infty 2} \end{bmatrix} = \begin{bmatrix} Y_{\infty 1} \\ Y_{\infty 2} \end{bmatrix} T_Y, \quad \operatorname{Re} \lambda_i(T_Y) \leqq 0 \quad \forall i,$$

(5.7) $$Y'_{\infty 1} Y_{\infty 2} = Y'_{\infty 2} Y_{\infty 1}$$

where

(5.8) $$J_\infty = \begin{bmatrix} (A - B_1 D'_{21} C_2)' & \gamma^{-2} C'_1 C_1 - C'_2 C_2 \\ -B_1 \tilde{D}'_\perp \tilde{D}_\perp B'_1 & -(A - B_1 D'_{21} C_2) \end{bmatrix},$$

and

(iii)

(5.9) $$\begin{bmatrix} X'_{\infty 2} X_{\infty 1} & \gamma^{-1} X'_{\infty 2} Y'_{\infty 2} \\ \gamma^{-1} Y'_{\infty 2} X_{\infty 2} & Y'_{\infty 2} Y_{\infty 1} \end{bmatrix} \geqq 0.$$

Remark 5.1 (connections with previous results). In the case where $X_{\infty 1}$ and $Y_{\infty 1}$ are invertible we have $(5.9) \geqq 0$ if and only if

$$\begin{bmatrix} (X'_{\infty 1})^{-1} & 0 \\ 0 & (Y'_{\infty 1})^{-1} \end{bmatrix} (5.9) \begin{bmatrix} X_{\infty 1}^{-1} & 0 \\ 0 & Y_{\infty 1}^{-1} \end{bmatrix} = \begin{bmatrix} X_\infty & \gamma^{-1} X_\infty Y_\infty \\ \gamma^{-1} Y_\infty X_\infty & Y_\infty \end{bmatrix} \geqq 0$$

$$\Rightarrow \begin{bmatrix} I & 0 \\ \gamma^{-1} Y_\infty & I \end{bmatrix} \begin{bmatrix} X_\infty & 0 \\ 0 & Y_\infty(I - \gamma^{-2} X_\infty Y_\infty) \end{bmatrix} \begin{bmatrix} I & \gamma^{-1} Y_\infty \\ 0 & I \end{bmatrix} \geqq 0.$$

Thus (5.9) is equivalent to the three conditions (1) $X_\infty \geqq 0$, (2) $Y_\infty \geqq 0$, and (3) $\rho(X_\infty Y_\infty) \leqq \gamma^2$. These last three conditions are given in [11] and [20] in the suboptimal case. The optimal cases in which X_∞ and Y_∞ exist were considered in [33] where a connection with vector interpolation is given.

The conditions of Theorem 5.1 treat the cases in which $X_{\infty 1}$ and/or $Y_{\infty 1}$ are singular. Examples of this type of optimality are given in [27].

THEOREM 5.2. *If the conditions of Theorem* 5.1 *are satisfied, then all internally stabilizing controllers K that satisfy* $\|F_l(P, K)\|_\infty \leqq \gamma$ *are given by*

(5.10) $K = F_l(\mathcal{K}, U)$ *with* $U \in \mathcal{RH}_+^\infty$, $\|U\|_\infty \leqq \gamma$, $\det(I - \mathcal{K}_{22}(\infty) U(\infty)) \neq 0$

where

(5.11) $$\mathcal{K}(s) = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} + \begin{bmatrix} C_{k1} \\ C_{k2} \end{bmatrix} (sE_k - A_k)^{\#} [B_{k1} \quad B_{k2}],$$

(5.12) $$E_k = Y'_{\infty 1} X_{\infty 1} - \gamma^{-2} Y'_{\infty 2} X_{\infty 2},$$

(5.13) $$B_{k1} = Y'_{\infty 1} B_1 D'_{21} + Y'_{\infty 2} C'_2,$$

(5.14) $$B_{k2} = Y'_{\infty 1} B_2 + \gamma^{-2} Y'_{\infty 2} C'_1 D_{12},$$

(5.15) $$C_{k1} = -D'_{12} C_1 X_{\infty 1} - B'_2 X_{\infty 2},$$

(5.16) $$C_{k2} = -C_2 X_{\infty 1} - \gamma^{-2} D_{21} B'_1 X_{\infty 2},$$

(5.17) $$A_k = E_k T_x + B_{k1} C_{k2}$$

(5.18) $$= T'_Y E_k + B_{k2} C_{k1}.$$ □

Remark 5.2 (computations, degree reduction, and the effective dimension of the free parameter). There are two possible consequences of E_k being singular. First, sE_k-A_k may have eigenvalues at infinity which do not appear as poles of $\mathscr{H}(s)$, and second, sE_k-A_k may be singular for all values of s. The same remarks applied to the realization of $Q_{aa}(s)$ given in § 4, and this was shown to be reducible to a standard state-space realization. The reason was twofold: First, $E_{aa}=R_{aa}+Q_{aa}$ satisfies the descriptor all-pass equations of Theorem 4.1, and second the realization of Q_{aa} has a particular structure as shown in § 4.3. It is easy to use the linear fractional relationship between $\mathscr{H}(s)$ and $Q_{aa}(s)$ to verify that the calculations, which were previously applied to $Q_{aa}(s)$ in § 4, are applicable to the realization of $\mathscr{H}(s)$, and these are now outlined. Suppose orthogonal changes of basis U and V are chosen so that $UE_kV=[\begin{smallmatrix}\hat{E}_k & 0\\ 0 & 0\end{smallmatrix}]$ in which $\hat{E}_k$ is nonsingular. Then (5.13) to (5.17) become

$$\hat{A}_k=UA_kV=\begin{bmatrix}\hat{A}_{k11} & \hat{A}_{k12}\\ \hat{A}_{k21} & \hat{A}_{k22}\end{bmatrix}, \tag{5.19}$$

$$\hat{B}_k=U[B_{k1}\quad B_{k2}]=\begin{bmatrix}\hat{B}_{k11} & \hat{B}_{k12}\\ \hat{B}_{k21} & \hat{B}_{k22}\end{bmatrix}, \tag{5.20}$$

$$\hat{C}_k=\begin{bmatrix}C_{k1}\\ C_{k2}\end{bmatrix}V=\begin{bmatrix}\hat{C}_{k11} & \hat{C}_{k12}\\ \hat{C}_{k21} & \hat{C}_{k22}\end{bmatrix} \tag{5.21}$$

in which the partitioning is compatible with that of $\hat{E}_k$. As explained in § 4, the state dimension of $\mathscr{H}(s)$ may be reduced by the rank defect of E_k by a singular perturbation type procedure. Direct calculation shows that a state-space model that is free of infinite eigenvalues is given by

$$A_{kr}=\hat{E}_k^{-1}[I\quad -\hat{A}_{k12}\hat{A}_{k22}^{\dagger}]\hat{A}_k\begin{bmatrix}I\\ 0\end{bmatrix}, \tag{5.22}$$

$$[B_{kr1}\quad B_{kr2}]=\hat{E}_k^{-1}[I\quad -\hat{A}_{k12}\hat{A}_{k22}^{\dagger}]\hat{B}_k, \tag{5.23}$$

$$\begin{bmatrix}C_{kr1}\\ C_{kr2}\end{bmatrix}=\hat{C}_k\begin{bmatrix}I\\ -\hat{A}_{k22}^{\dagger}\hat{A}_{k21}\end{bmatrix}, \tag{5.24}$$

$$\begin{bmatrix}D_{kr11} & D_{kr12}\\ D_{kr21} & D_{kr22}\end{bmatrix}=\begin{bmatrix}0 & I\\ I & 0\end{bmatrix}-\begin{bmatrix}\hat{C}_{k12}\\ \hat{C}_{k22}\end{bmatrix}\hat{A}_{k22}^{\dagger}[\hat{B}_{k21}\quad \hat{B}_{k22}]. \tag{5.25}$$

To explicitly show the reduction in dimension of the free parameter $U(s)$, we select orthogonal matrices Y and Z (which always exist) such that

$$Y[C_{kr2}\quad D_{kr21}]=\begin{bmatrix}\bar{C}_{kr2} & \bar{D}_{kr21}\\ 0 & 0\end{bmatrix}, \tag{5.26}$$

$$\begin{bmatrix}B_{kr2}\\ D_{kr12}\end{bmatrix}Z=\begin{bmatrix}\bar{B}_{kr2} & 0\\ \bar{D}_{kr12} & 0\end{bmatrix}, \tag{5.27}$$

$$YD_{kr22}Z=\begin{bmatrix}\bar{D}_{kr22} & 0\\ 0 & \gamma^{-1}I\end{bmatrix}. \tag{5.28}$$

Thus

$$\mathscr{H}(s)=\begin{bmatrix}D_{kr11} & \bar{D}_{kr12}\\ \bar{D}_{kr21} & \bar{D}_{kr22}\end{bmatrix}+\begin{bmatrix}C_{kr1}\\ \bar{C}_{kr2}\end{bmatrix}(sI-A_{kr})^{-1}[B_{kr1}\quad \bar{B}_{kr2}]. \qquad \square \tag{5.29}$$

5.1. A review of controller parameterization theory. The purpose of this section is to show how the original H^∞ control problem may be recast as a four block general distance problem. Our treatment of parameterization theory will be brief, as this material is standard and the details already appear in several places [10], [14], [15], [43]. Since P_{11}, P_{12}, P_{21}, and P_{22} share the same state-space it is clear the K stabilizes $F_l(P, K)$ if and only if it stabilizes P_{22}. Since (A, B_2, C_2) is assumed stabilizable and detectable, such controllers always exist. Let

$$P_{22} = N_r D_r^{-1} = D_l^{-1} N_l \tag{5.30}$$

be left and right coprime stable rational matrix fraction descriptions of P_{22} and

$$\begin{bmatrix} V_r & U_r \\ -N_l & D_l \end{bmatrix} \begin{bmatrix} D_r & -U_l \\ N_r & V_l \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} \tag{5.31}$$

the corresponding Bezout identities. All the matrices in (5.31) belong to $\mathcal{H}_+^\infty$, and the set of all compensators which stabilize P_{22} and thus also P are given by [43]

$$K = F_l(K_0, Q), \qquad Q \in \mathcal{H}_+^\infty \tag{5.32}$$

where

$$K_0(s) = \begin{bmatrix} -V_r^{-1} U_r & -V_r^{-1} \\ V_l^{-1} & V_l^{-1} N_r \end{bmatrix}. \tag{5.33}$$

Since

$$K(I - P_{22}K)^{-1} = (-D_r Q - U_l) D_l$$

we obtain

$$\begin{aligned} \mathcal{R}(s) &= P_{11} + P_{12}K(I - P_{22}K)^{-1}P_{21} \\ &= (P_{11} - P_{12}U_l D_l P_{21}) - (P_{12}D_r)Q(D_l P_{21}) \\ &= T_{11} + T_{12}QT_{21}, \end{aligned} \tag{5.34}$$

which is an affine parameterization of all internally stable closed loops.

Since (A, B_2) is stabilizable there exists a state feedback matrix F such that $A - B_2F$ is stable. Similarly, since (C_2, A) is detectable there exists an output injection matrix H such that $A - HC_2$ is stable. Given any such pair of stabilizing matrices F and H, the right and left coprime factorizations of P_{22} together with the solutions of the Bezout identities are given by [10], [14], [15], and [43]

$$\begin{bmatrix} D_r & -U_l \\ N_r & V_l \end{bmatrix} \overset{s}{=} \left[\begin{array}{c|cc} A - B_2F & B_2 & H \\ \hline -F & I & 0 \\ C_2 & 0 & I \end{array}\right] \tag{5.35}$$

and

$$\begin{bmatrix} V_r & U_r \\ -N_l & D_l \end{bmatrix} \overset{s}{=} \left[\begin{array}{c|cc} A - HC_2 & B_2 & H \\ \hline F & I & 0 \\ -C_2 & 0 & I \end{array}\right]. \tag{5.36}$$

Substituting (5.35) and (5.36) into (5.33) yields

$$K_0(s) \overset{s}{=} \left[\begin{array}{c|cc} A - B_2F - HC_2 & H & B_2 \\ \hline -F & 0 & I \\ -C_2 & I & 0 \end{array}\right] \tag{5.37}$$

after simplification. We will now make the following specific choices of the stabilizing matrices F and H which will lead to all-pass properties in T_{12} and T_{21} in (5.34). Specifically, we define F and H by

$$F = D_{12}'C_1 + B_2'X, \tag{5.38}$$

$$H = B_1D_{21}' + YC_2' \tag{5.39}$$

where X and Y are the unique stabilizing solutions to

$$X(A - B_2D_{12}'C_1) + (A - B_2D_{12}'C_1)'X - XB_2B_2'X + C_1'D_\perp D_\perp' C_1 = 0 \tag{5.40}$$

and

$$Y(A - B_1D_{21}'C_2)' + (A - B_1D_{21}'C_2)Y - YC_2'C_2Y + B_1\tilde{D}_\perp'\tilde{D}_\perp B_1' = 0. \tag{5.41}$$

Direct substitution into (5.34) leads to

$$\begin{bmatrix} T_{11} & T_\perp & T_{12} \\ \tilde{T}_\perp & 0 & 0 \\ T_{21} & 0 & 0 \end{bmatrix}(s) \overset{s}{=} \left[\begin{array}{cc|ccc} A - B_2F & B_2F & B_1 & -X^\dagger C_1'D_\perp & B_2 \\ 0 & A - HC_2 & B_1 - HD_{21} & 0 & 0 \\ \hline C_1 - D_{12}F & D_{12}F & 0 & D_\perp & D_{12} \\ 0 & -\tilde{D}_\perp B_1'Y^\dagger & \tilde{D}_\perp & 0 & 0 \\ 0 & C_2 & D_{21} & 0 & 0 \end{array}\right]. \tag{5.42}$$

With this particular choice of F and H, $[T_\perp \mid T_{12}]$ and $[\tilde{T}_\perp^\sim \mid T_{21}^\sim]$ are all-pass [10], [14], [15], [43]. We call $T_\perp(s)$ and $\tilde{T}_\perp(s)$ all-pass extensions of T_{12} and T_{21}, respectively, and this all-pass property allows us to write

$$\left\| T_{11} - [T_\perp \ \ T_{12}] \begin{bmatrix} 0 & 0 \\ 0 & Q \end{bmatrix} \begin{bmatrix} \tilde{T}_\perp \\ T_{21} \end{bmatrix} \right\|_\infty = \left\| \begin{matrix} R_{11} & R_{12} \\ R_{21} & R_{22} - Q \end{matrix} \right\|_\infty \tag{5.43}$$

where

$$R(s) = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} = \begin{bmatrix} T_\perp^\sim \\ T_{12}^\sim \end{bmatrix} T_{11}[\tilde{T}_\perp^\sim \ \ T_{21}^\sim]. \tag{5.44}$$

Substituting (5.42) into (5.44) gives

$$R(s) \overset{s}{=} \left[\begin{array}{cc|cc} -(A - B_2F)' & XB_1(B_1 - HD_{21})' & XB_1\tilde{D}_\perp' & XB_1D_{21}' \\ 0 & -(A - HC_2)' & Y^\dagger B_1\tilde{D}_\perp' & -C_2' \\ \hline D_\perp' C_1X^\dagger & 0 & 0 & 0 \\ -B_2' & FY & 0 & 0 \end{array}\right], \tag{5.45}$$

which describes the four block problem to be solved.

In the case where X and/or Y are singular, this realization will not be minimal and it will be convenient to delete the nonminimal states as follows [31], [32]. Let

$$X = U\begin{bmatrix} X_1 & 0 \\ 0 & 0 \end{bmatrix}U', \quad X_1 > 0 \quad \text{where } U = [U_1 \ \ U_2] \text{ is unitary;}$$

$$Y = V\begin{bmatrix} Y_1 & 0 \\ 0 & 0 \end{bmatrix}V', \quad Y_1 > 0 \quad \text{where } V = [V_1 \ \ V_2] \text{ is unitary.}$$

The Riccati equations for X and Y then give that

$$U_1'(A - B_2D_{12}'C_1)U_2 = U_1'(A - B_2F)U_2 = 0, \qquad D_\perp' C_1U_2 = 0$$

$$V_2'(A - B_1D_{21}'C_2)V_1 = V_2'(A - HC_2)V_1 = 0, \qquad V_2'B_1\tilde{D}_\perp' = 0$$

$$B_1 - HD_{21} = B_1\tilde{D}_\perp'\tilde{D}_\perp - YC_2'D_{21} \Rightarrow V_2'(B_1 - HD_{21}) = 0$$

$$C_1 - D_{12}F = D_\perp D_\perp' C_1 - D_{12}B_2'X \Rightarrow (C_1 - D_{12}F)U_2 = 0$$

The state transformation $[\begin{smallmatrix} U' & 0 \\ 0 & V' \end{smallmatrix}]$ applied to $R(s)$ then exhibits the nonminimal modes which may be removed, and this followed by the state transformation $[\begin{smallmatrix} X_1^{-1} & 0 \\ 0 & Y_1 \end{smallmatrix}]$ gives

$$(5.46)\quad R(s)=\left[\begin{array}{c|cc} A_R & B_{R1} & B_{R2} \\ \hline C_{R1} & 0 & 0 \\ C_{R2} & 0 & 0 \end{array}\right] \overset{s}{=} \left[\begin{array}{cc|cc} -A_F' & U_1'B_1B_{1H}' & U_1'B_1\tilde{D}_\perp' & U_1'B_1D_{21}' \\ 0 & -A_H' & V_1'B_1\tilde{D}_\perp' & -Y_1V_1'C_2' \\ \hline D_\perp'C_1U_1 & 0 & 0 & 0 \\ -B_2'U_1X_1 & FV_1 & 0 & 0 \end{array}\right].$$

An alternative realization for $R(s)$ can be obtained via a state-space transformation

$$T_R=\begin{bmatrix} I & -U_1'V_1 \\ 0 & I \end{bmatrix}.$$

That is,

$$(5.47)\qquad R(s)\overset{s}{=}\left[\begin{array}{cc|cc} -A_F' & C_{1F}'C_1V_1 & 0 & U_1'H \\ 0 & -A_H' & V_1'B_1\tilde{D}_\perp' & -Y_1V_1'C_2' \\ \hline D_\perp'C_1U_1 & D_\perp'C_1V_1 & 0 & 0 \\ -B_2'U_1X_1 & D_{12}'C_1V_1 & 0 & 0 \end{array}\right]$$

where

$$\begin{aligned} A_F &= X_1U_1'(A-B_2F)U_1X_1^{-1} \\ A_H &= Y_1^{-1}V_1'(A-HC_2)V_1Y_1 \\ B_{1H} &= Y_1^{-1}V_1'(B_1-HD_{21}) = Y_1^{-1}V_1'B_1\tilde{D}_\perp'\tilde{D}_\perp - V_1'C_2'D_{21} \\ C_{1F} &= (C_1-D_{12}F)U_1X_1^{-1} = D_\perp D_\perp'C_1U_1X_1^{-1} - D_{12}B_2'U_1 \end{aligned}$$

and the Riccati equations for X and Y can be rewritten as

$$(5.48)\qquad A_F'X_1^{-1}+X_1^{-1}A_F+C_{1F}'C_{1F}=0,$$

$$(5.49)\qquad A_HY_1^{-1}+Y_1^{-1}A_H'+B_{1H}B_{1H}'=0.$$

5.2. Necessary and sufficient conditions. A necessary condition for $\left\|R+\begin{bmatrix} 0 & 0 \\ 0 & Q \end{bmatrix}\right\|_\infty \leqq \gamma$ is $\|R_{11} \quad R_{12}\|_\infty \leqq \gamma \Leftrightarrow \|T_\perp^\sim T_{11}[\tilde{T}_\perp^\sim \quad T_{21}^\sim]\|_\infty \leqq \gamma \Leftrightarrow \|T_\perp^\sim T_{11}\|_\infty \leqq \gamma$.

However,

$$\begin{aligned} T_\perp^\sim T_{11} &\overset{s}{=} \left[\begin{array}{c|c} -(A-B_2F)' & XB_1 \\ \hline D_\perp'C_1X^\dagger & 0 \end{array}\right] \\ &\overset{s}{=} \left[\begin{array}{c|c} -A_F' & U_1'B_1 \\ \hline D_\perp'C_1U_1 & 0 \end{array}\right] \end{aligned}$$

so that the corresponding spectral factorization Riccati equation to find R_{13} (note that $R_{13}R_{13}^\sim=\gamma^2I-[R_{11} \quad R_{12}][R_{11}^\sim \quad R_{12}^\sim]^\sim=\gamma^2I-T_\perp^\sim T_{11}T_{11}^\sim T_\perp$) has a solution $\hat{X}\geqq 0$ satisfying

$$H_{\hat{X}}\begin{bmatrix} I \\ \hat{X} \end{bmatrix}=\begin{bmatrix} I \\ \hat{X} \end{bmatrix}T_{\hat{X}}, \qquad \operatorname{Re}\lambda_i(T_{\hat{X}})\leqq 0 \quad \forall i$$

where

$$(5.50)\qquad H_{\hat{X}}=\begin{bmatrix} A_F & \gamma^{-2}U_1'C_1'D_\perp D_\perp'C_1U_1 \\ -U_1'B_1B_1'U_1 & -A_F' \end{bmatrix}.$$

Similarly, the condition $\|R_{\widetilde{11}} \quad R_{\widetilde{21}}\|_\infty \leqq \gamma$ implies that there exists $\hat{Y} \geqq 0$ such that

$$H_{\hat{Y}} \begin{bmatrix} I \\ \hat{Y} \end{bmatrix} = \begin{bmatrix} I \\ \hat{Y} \end{bmatrix} T_{\hat{Y}}, \qquad \operatorname{Re} \lambda_i(T_{\hat{Y}}) \leqq 0 \quad \forall i,$$

where

$$H_{\hat{Y}} = \begin{bmatrix} A_H' & \gamma^{-2} V_1' B_1 \tilde{D}_\perp' \tilde{D}_\perp B_1' V_1 \\ -V_1' C_1' C_1 V_1 & -A_H \end{bmatrix}. \tag{5.51}$$

In order to apply the necessary and sufficient conditions of Theorem 4.2, the solutions to the Riccati equations (3.51) and (3.52) (which determine R_{13} and R_{31}) in $R(s)$ are required, and they are now given.

LEMMA 5.3. *For the realization of $R(s)$ defined in* (5.46) *the Riccati equations*:

$$-P_1 A_R' - A_R P_1 + B_{R1} B_{R1}' + B_{R2} B_{R2}' + \gamma^{-2} P_1 C_{R1}' C_{R1} P_1 = 0 \tag{5.52a}$$

$$-Q_1 A_R - A_R' Q_1 + C_{R1}' C_{R1} + C_{R2}' C_{R2} + \gamma^{-2} Q_1 B_{R1}' B_{R1} Q_1 = 0 \tag{5.52b}$$

are satisfied for

$$P_1 = \begin{bmatrix} \hat{X} & 0 \\ 0 & Y_1 \end{bmatrix}, \qquad Q_1 = T_R' Q_0 T_R$$

$$T_R = \begin{bmatrix} I & -U_1' V_1 \\ 0 & I \end{bmatrix}, \qquad Q_0 = \begin{bmatrix} X_1 & 0 \\ 0 & \hat{Y} \end{bmatrix},$$

and

$$\operatorname{Re} \lambda_i(A_R - \gamma^{-2} P_1 C_{R1}' C_{R1}) \geqq 0 \quad \forall i,$$

$$\operatorname{Re} \lambda_i(A_R - \gamma^{-2} B_{R1} B_{R1}' Q_1) \geqq 0 \quad \forall i.$$

Proof. P_1 is simply shown to satisfy (5.52) by (5.49) and (5.50) with antistability given by the stability of $T_{\hat{X}}$ and A_H. An analogous derivation applies to Q_0 for the alternative realization of R in (5.47) and then $T_R'^{-1}$ relates Q_1 to Q_0. □

Proof of Theorem 5.1. Theorem 4.2 and the above calculations give the necessary and sufficient conditions for controllers to exist to be that $\hat{X} \geqq 0$ and $\hat{Y} \geqq 0$ exist, and

$$\lambda_{\max}(P_1 Q_1) \leqq \gamma^2.$$

To write this condition as that stated in the theorem, we write

$$\begin{aligned} \gamma^2 I - P_1 Q_1 &= (\gamma^2 T_R^{-1} - P_1 T_R' Q_0) T_R \\ &= \begin{bmatrix} \gamma I & 0 \\ 0 & Y_1 \end{bmatrix} \Pi(\gamma) \begin{bmatrix} X_1 & 0 \\ 0 & \gamma I \end{bmatrix} T_R \end{aligned} \tag{5.53}$$

in which

$$\Pi(\gamma) = \begin{bmatrix} \gamma X_1^{-1} - \gamma^{-1} \hat{X} & U_1' V_1 \\ V_1' U_1 & \gamma Y_1^{-1} - \gamma^{-1} \hat{Y} \end{bmatrix}.$$

Since P_1 and Q_1 are both monotonically decreasing functions of γ [51], $\lambda_{\max}(P_1 Q_1)$ is also a monotonically decreasing function of γ, and so $\lambda_{\max}(P_1 Q_1) \leqq \gamma^2$ if and only if $\Pi(\gamma) \geqq 0$.

This condition now needs to be written in terms of $X_{\infty 1}$, $X_{\infty 2}$, $Y_{\infty 1}$, and $Y_{\infty 2}$. The Riccati equations for X_1 (5.48) give, as in [11, § VII.C], that

$$\begin{bmatrix} X_1^{-1} & -\gamma^{-2}I \\ I & 0 \end{bmatrix} H_{\hat{X}} \begin{bmatrix} 0 & I \\ -\gamma^{-2}I & \gamma^2 X_1^{-1} \end{bmatrix} = \begin{bmatrix} U_1' & 0 \\ 0 & U_1' \end{bmatrix} H_\infty \begin{bmatrix} U_1 & 0 \\ 0 & U_1 \end{bmatrix}.$$

Hence

$$\begin{bmatrix} U_1' & 0 \\ O & U_1' \end{bmatrix} H_\infty \begin{bmatrix} U_1 & 0 \\ 0 & U_1 \end{bmatrix} \begin{bmatrix} X_1^{-1} & -\gamma^{-2}I \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ \hat{X} \end{bmatrix} = \begin{bmatrix} X_1^{-1} & -\gamma^{-2}I \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ \hat{X} \end{bmatrix} T_{\hat{X}}$$

$$= \begin{bmatrix} X_1^{-1} - \gamma^{-2}\hat{X} \\ I \end{bmatrix} T_{\hat{X}}$$

and exploiting the structure of $U'(A - B_2 D_{12}' C_1)U$ and $U' C_1' D_\perp$ give

$$H_\infty \begin{bmatrix} X_{\infty 1} \\ X_{\infty 2} \end{bmatrix} = \begin{bmatrix} X_{\infty 1} \\ X_{\infty 2} \end{bmatrix} T_x$$

where

$$X_{\infty 1} = U_1(X_1^{-1} - \gamma^{-2}\hat{X})U_1' + U_2 U_2',$$
$$X_{\infty 2} = U_1 U_1'$$

satisfy

$$X_{\infty 1}' X_{\infty 2} = X_{\infty 2}' X_{\infty 1}$$

and all such matrices are given by $\left[\begin{smallmatrix} X_{\infty 1} \\ X_{\infty 2} \end{smallmatrix}\right] S$ for a nonsingular S. (This comes from the strict stability of $U_2'(A - B_2 D_{12}' C_1)U_2$ and the uniqueness of $\hat{X}$.) Furthermore, exploiting the structure of $\left[\begin{smallmatrix} U' & 0 \\ 0 & U' \end{smallmatrix}\right] H_\infty \left[\begin{smallmatrix} U & 0 \\ 0 & U \end{smallmatrix}\right]$ shows that if $X_{\infty 1}' X_{\infty 2} = X_{\infty 2}' X_{\infty 1}$ then $\hat{X} = \hat{X}'$ exists. An analogous argument applies to $Y_{\infty 1}$, $Y_{\infty 2}$, and $\hat{Y}$. The final condition $\Pi(\gamma) \geqq 0$ is equivalent to

$$0 \leqq \begin{bmatrix} U_1 & 0 \\ 0 & V_1 \end{bmatrix} \Pi(\gamma) \begin{bmatrix} U_1' & 0 \\ 0 & V_1' \end{bmatrix}$$
$$= \gamma \begin{bmatrix} X_{\infty 2}' X_{\infty 1} & \gamma^{-1} X_{\infty 2}' Y_{\infty 2} \\ \gamma^{-1} Y_{\infty 2}' X_{\infty 2} & Y_{\infty 2}' Y_{\infty 1} \end{bmatrix}. \qquad \square$$

5.3. Characterization of all solutions. Theorem 4.4 characterizes all optimal solutions to the model matching problem and Theorem 5.2 claims to do the same for the feedback control problem. A derivation of the formula in Theorem 5.2 can be obtained by applying Theorem 4.4 to generate all $Q(s)$'s, and then substituting into $K_0(s)$ to generate all $K(s)$'s. Finally, all the nonminimal modes are removed from this parameterization. In the rest of this section we assume, without loss of generality, that $\gamma = 1$ for simplicity.

Applying Theorem 4.2 to the realization of R given in (5.46) gives

$$\begin{bmatrix} Q_{22} & Q_{24} \\ Q_{42} & Q_{44} \end{bmatrix} = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} + \begin{bmatrix} C_{R2} P_1 \\ -B_{R2}' \end{bmatrix}$$
$$\cdot ((I - Q_1 P_1)s - (-A_R' - Q_1 A_R P_1 + Q_1 B_{R1} B_{R1}' + C_{R1}' C_{R1} P_1))^{-1}$$
$$\cdot [Q_1 B_{R2} \quad -C_{R2}'].$$

Recall from (5.53) that

$$(I - Q_1 P_1) = T_R' \begin{bmatrix} X_1 & 0 \\ 0 & I \end{bmatrix} \Pi \begin{bmatrix} I & 0 \\ 0 & Y_1 \end{bmatrix}$$

in which $\Pi = \Pi(1)$. Changing coordinates gives

$$\begin{bmatrix} Q_{22} & Q_{24} \\ Q_{42} & Q_{44} \end{bmatrix} = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} + \begin{bmatrix} \hat{C}_{01} \\ \hat{C}_{02} \end{bmatrix} (s\Pi - \hat{A}_0)^{-1} [\hat{B}_{01} \quad \hat{B}_{02}]$$

$$\text{(5.54a)} \quad \hat{A}_0 = \begin{bmatrix} X_1^{-1} & 0 \\ 0 & I \end{bmatrix} (-T_R'^{-1} A_R' - Q_0 T_R A_R P_1 + Q_0 T_R B_{R1} B'_{R1} + T_R'^{-1} C'_{R1} C_{R1} P_1) \begin{bmatrix} I & 0 \\ 0 & Y_1^{-1} \end{bmatrix},$$

$$\text{(5.55)} \quad \begin{bmatrix} \hat{C}_{01} \\ \hat{C}_{02} \end{bmatrix} = \begin{bmatrix} C_{R2} P_1 \\ -B'_{R2} \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & Y_1^{-1} \end{bmatrix} = \begin{bmatrix} -B_2' U_1 X_1 \hat{X} & F V_1 \\ -D_{21} B_1' U_1 & C_2 V_1 \end{bmatrix},$$

$$\text{(5.56)} \quad \begin{aligned} [\hat{B}_{01} \quad \hat{B}_{02}] &= \begin{bmatrix} X_1^{-1} & 0 \\ 0 & I \end{bmatrix} [Q_0 T_R B_{R2} \quad -T_R'^{-1} C'_{R2}] \\ &= \begin{bmatrix} U_1' H & U_1' B_2 \\ -\hat{Y} Y_1 V_1' C_2' & -V_1' C_1' D_{12} \end{bmatrix}, \end{aligned}$$

$$\text{(5.54b)} \quad \begin{aligned} \hat{A}_{011} &= X_1^{-1} A_F + A_F' \hat{X} + X_1^{-1} U_1' C_1' D_\perp D_\perp' C_1 U_1 \hat{X} \\ &= X_1^{-1} A_F - X_1^{-1} A_F X_1 \hat{X} - U_1' B_2 B_2' U_1 X_1 \hat{X} \end{aligned}$$

$$\text{(5.54c)} \quad \begin{aligned} \hat{A}_{012} &= -(U_1' B_1 B'_{1H} + U_1' V_1 A_H') = -U_1'(B_1(B_1 - HD_{21})' + Y(A - HC_2)') V_1 Y_1^{-1} \\ &= U_1' A Y V_1 Y_1^{-1} = U_1' A V_1, \end{aligned}$$

$$\text{(5.54d)} \quad \begin{aligned} \hat{A}_{021} &= -(B_{1H} B_1' U_1 - V_1' U_1 A_F) + \hat{Y} V_1' B_1 \tilde{D}_\perp' \tilde{D}_\perp B_1' U_1 + V_1' C_1' D_\perp D_\perp' C_1 U_1 \hat{X} \\ &= (\hat{Y} - Y_1^{-1}) V_1' B_1 \tilde{D}_\perp' \tilde{D}_\perp B_1' U_1 + V_1' C_1' D_\perp D_\perp' C_1 U_1 (\hat{X} - X_1^{-1}) \\ &\quad - V_1'(A - B_2 D_{12}' C_1 - B_1 D_{21}' C_2)' U_1, \end{aligned}$$

$$\text{(5.54e)} \quad \begin{aligned} \hat{A}_{022} &= A_H Y_1^{-1} + \hat{Y} A_H' + \hat{Y} V_1' B_1 \tilde{D}_\perp' \tilde{D}_\perp B_1' V_1 Y_1^{-1} \\ &= A_H Y_1^{-1} + \hat{Y} Y_1 (-A_H Y_1^{-1} - V_1' C_2' C_2 V_1) \\ &= (I - \hat{Y} Y_1) A_H Y_1^{-1} - \hat{Y} Y_1 V_1' C_2' C_2 V_1. \end{aligned}$$

The characterization of all $K(s)$ is given by

$$\begin{aligned} K(s) &= F_l\left(K_0, F_l\left(\begin{bmatrix} Q_{22} & Q_{24} \\ Q_{42} & Q_{44} \end{bmatrix}, U\right)\right) \\ &= F_l(\mathcal{K}, U) \end{aligned}$$

where

$$\mathcal{K} = F_l \left\{ \begin{bmatrix} K_{011} & 0 & K_{012} & 0 \\ 0 & 0 & 0 & I \\ K_{021} & 0 & K_{022} & 0 \\ 0 & I & 0 & 0 \end{bmatrix}, \begin{bmatrix} Q_{22} & Q_{24} \\ Q_{42} & Q_{44} \end{bmatrix} \right\}.$$

Substituting for the realization gives that

$$\text{(5.57)} \quad \mathcal{K}(s) = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} + \hat{C}_1 (s\hat{E}_1 - \hat{A}_1)^{-1} \hat{B}_1$$

where

$$\hat{E}_1 = \begin{bmatrix} I & 0 \\ 0 & \Pi \end{bmatrix}, \tag{5.58}$$

$$\hat{A}_1 = \begin{bmatrix} A - B_2F - HC_2 & B_2\hat{C}_{01} \\ -\hat{B}_{01}C_2 & \hat{A}_0 \end{bmatrix}, \tag{5.59}$$

$$\hat{B}_1 = \begin{bmatrix} H & B_2 \\ \hat{B}_{01} & \hat{B}_{02} \end{bmatrix}, \tag{5.60}$$

$$\hat{C}_1 = \begin{bmatrix} -F & \hat{C}_{01} \\ -C_2 & \hat{C}_{02} \end{bmatrix}. \tag{5.61}$$

This realization contains rank (X) uncontrollable modes and rank (Y) unobservable modes and these are exhibited by the following transformation. First, define

$$X_{\infty 11} = X_1^{-1} - \hat{X}, \qquad Y_{\infty 11} = Y_1^{-1} - \hat{Y}$$

$$T_l = \begin{bmatrix} V' & 0 & 0 \\ -U_1' & I & 0 \\ -Y_{\infty 11}'V_1' & 0 & I \end{bmatrix}, \qquad T_r = \begin{bmatrix} U & U_1X_{\infty 11} & V_1 \\ 0 & I & 0 \\ 0 & 0 & I \end{bmatrix}, \tag{5.62}$$

$$\hat{E}_2 = T_l\hat{E}_1T_r = \begin{bmatrix} V'U & V'U_1X_{\infty 11} & \begin{bmatrix} I \\ 0 \end{bmatrix} \\ [-I \quad 0] & 0 & 0 \\ -Y_{\infty 11}V_1'U & V_1'U_1 - Y_{\infty 11}V_1'U_1X_{\infty 11} & 0 \end{bmatrix}, \tag{5.63}$$

$$\hat{A}_2 = T_l\hat{A}_1T_r = \begin{bmatrix} \hat{A}_{211} & \hat{A}_{212} & \hat{A}_{213} \\ \hat{A}_{221} & \hat{A}_{222} & \hat{A}_{223} \\ \hat{A}_{231} & \hat{A}_{232} & \hat{A}_{233} \end{bmatrix}, \tag{5.64}$$

$$\hat{B}_2 = T_l\hat{B}_1 = \begin{bmatrix} \hat{B}_{21} \\ \hat{B}_{22} \\ \hat{B}_{23} \end{bmatrix}, \qquad \hat{C}_2 = \hat{C}_1T_r = [\hat{C}_{21} \quad \hat{C}_{22} \quad \hat{C}_{23}], \tag{5.65}$$

$$\hat{B}_{22} = 0 \quad \text{by (5.56)},$$

$$\hat{C}_{23} = 0 \quad \text{by (5.57)},$$

$$\begin{aligned} [\hat{A}_{221} \quad \hat{A}_{222} \quad \hat{A}_{223}] &= -U_1'[A - B_2F - HC_2 \quad -B_2B_2'U_1X_1\hat{X} \quad B_2FV_1]T_r \\ &\quad + [-U_1'HC_2 \quad \hat{A}_{011} \quad \hat{A}_{012}]T_r \\ &= [-U_1'(A - B_2F) \quad X_1^{-1}A_FX_1X_{\infty 11} \quad \hat{A}_{012} - U_1'B_2FV_1]T_r \\ \Rightarrow \hat{A}_{221} &= [-U_1'(A - B_2F)U_1 \quad 0], \qquad \hat{A}_{222} = 0, \qquad \hat{A}_{223} = 0 \end{aligned}$$

Hence the first rank (X) states are uncontrollable. Furthermore,

$$\hat{A}_{213} = V'(A - B_2F - HC_2)V_1 + V'B_2FV_1 = \begin{bmatrix} V_1'(A - HC_2)V_1 \\ 0 \end{bmatrix},$$

$$\begin{aligned} [\hat{A}_{231} \quad \hat{A}_{232} \quad \hat{A}_{233}] &= -Y_{\infty 11}V_1'[A - B_2F - HC_2 \quad -B_2B_2'U_1X_1\hat{X} \quad B_2FV_1]T_r \\ &\quad + [\hat{Y}Y_1V_1'C_2'C_2 \quad \hat{A}_{021} \quad \hat{A}_{022}]T_r \end{aligned}$$

$\Rightarrow \hat{A}_{233} = 0$ and the last rank (Y) states are unobservable. A reduced state model can hence be obtained and its realization transformed as follows:

$$\hat{E}_3 = [V_2 \quad -V_1]\begin{bmatrix} V_2'U_2 & V_2'U_1X_{\infty 11} \\ -Y_{\infty 11}V_1'U_2 & V_1'U_1 - Y_{\infty 11}V_1'U_1X_{\infty 11}\end{bmatrix}\begin{bmatrix} U_2' \\ U_1' \end{bmatrix} \tag{5.66}$$

without loss of generality (see proof of Theorem 5.1) we can assume

$$X_{\infty 1} = U\begin{bmatrix} X_{\infty 11} & 0 \\ 0 & I \end{bmatrix}U', \qquad X_{\infty 2} = U_1U_1',$$

$$Y_{\infty 1} = V\begin{bmatrix} Y_{\infty 11} & 0 \\ 0 & I \end{bmatrix}V', \qquad Y_{\infty 2} = V_1V_1'$$

and hence

$$\hat{E}_3 = Y_{\infty 1}'X_{\infty 1} - Y_{\infty 2}'X_{\infty 2} = E_k,$$

$$\begin{aligned}
\hat{B}_3 = [B_{k1} \quad B_{k2}] &= [V_2 \quad -V_1]\begin{bmatrix} V_2' & 0 \\ -Y_{\infty 11}'V_1' & I \end{bmatrix}\begin{bmatrix} H & B_2 \\ -\hat{Y}Y_1V_1'C_2' & -V_1'C_1'D_{12}\end{bmatrix} \\
&= [Y_{\infty 1}' \quad -V_1]\begin{bmatrix} H & B_2 \\ -\hat{Y}Y_1V_1'C_2' & -V_1'C_1'D_{12}\end{bmatrix} \\
&= [Y_{\infty 1}'(YC_2' + B_1D_{21}') + V_1(I - Y_{\infty 11}Y_1)V_1'C_2', \\
&\qquad Y_{\infty 1}'B_2 + V_1V_1'C_1'D_{12}] \\
&= [Y_{\infty 2}'C_2' + Y_{\infty 1}'B_1D_{21}' \quad Y_{\infty 1}'B_2 + Y_{\infty 2}'C_1'D_{12}],
\end{aligned} \tag{5.67}$$

$$\begin{aligned}
\hat{C}_3 = \begin{bmatrix} C_{k1} \\ C_{k2} \end{bmatrix} &= \begin{bmatrix} -F & -B_2'U_1X_1\hat{X} \\ -C_2 & -D_{21}B_1'U_1 \end{bmatrix}\begin{bmatrix} U_2 & U_1X_{\infty 11} \\ 0 & I \end{bmatrix}\begin{bmatrix} U_2' \\ U_1' \end{bmatrix} \\
&= -\begin{bmatrix} F & B_2'U_1X_1\hat{X} \\ C_2 & D_{21}B_1'U_1 \end{bmatrix}\begin{bmatrix} X_{\infty 1} \\ U_1' \end{bmatrix} \\
&= -\begin{bmatrix} D_{12}'C_1X_{\infty 1} + B_2'X_{\infty 2} \\ C_2X_{\infty 1} + D_{21}B_1'X_{\infty 2} \end{bmatrix},
\end{aligned} \tag{5.68}$$

$$\begin{aligned}
\hat{A}_3 &= [Y_{\infty 1}' \quad -V_1]\begin{bmatrix} A - B_2F - HC_2 & -B_2B_2'U_1X_1\hat{X} \\ \hat{Y}Y_1V_1'C_2'C_2 & \hat{A}_{021} \end{bmatrix}\begin{bmatrix} X_{\infty 1} \\ U_1' \end{bmatrix} \\
&= [Y_{\infty 1}' \quad -Y_{\infty 2}']\left\{H_\infty + \begin{bmatrix} -B_1D_{21}' \\ C_2' \end{bmatrix}[C_2 \quad D_{21}B_1'] + \begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}\right\}\begin{bmatrix} X_{\infty 1} \\ X_{\infty 2} \end{bmatrix}
\end{aligned} \tag{5.69}$$

where

$$L_{11} = -B_2B_2'X - YC_2'C_2,$$

$$L_{12} = B_2B_2'(I - U_1X_1\hat{X}U_1') - B_1\tilde{D}_\perp'\tilde{D}_\perp B_1',$$

$$L_{21} = -(I - V_1\hat{Y}Y_1V_1')C_2'C_2 + C_1'D_\perp D_\perp'C_1,$$

$$L_{22} = (A - B_2D_{12}'C_1 - B_1D_{21}'C_2)' + V_1\hat{A}_{021}U_1',$$

$$V_1'L_{22}U_1 = (\hat{Y} - Y_1^{-1})V_1'B_1\tilde{D}_\perp'\tilde{D}_\perp B_1'U_1 + V_1'C_1'D_\perp D_\perp'C_1U_1(\hat{X} - X_1^{-1}).$$

The equations $XX_{\infty 1}+U_1X_1\hat{X}U_1'=U_1U_1'$ and $Y_{\infty 1}Y+V_1\hat{Y}Y_1V_1'=V_1V_1'$ and direct substitution gives

$$[Y_{\infty 1} \quad -Y_{\infty 2}]\begin{bmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{bmatrix}\begin{bmatrix} X_{\infty 1} \\ X_{\infty 2} \end{bmatrix}=0$$

and finally $\hat{A}_3=A_k$ since

$$H_\infty\begin{bmatrix} X_{\infty 1} \\ X_{\infty 2} \end{bmatrix}=\begin{bmatrix} X_{\infty 1} \\ X_{\infty 2} \end{bmatrix}T_x. \qquad \square$$

5.4. Removing assumption (v). In this section we show how we might tackle problems in which $D_{11}\neq 0$ and/or $D_{22}\neq 0$. The idea is to show that assumption (v) results in no loss of generality, and that the original problem with $D_{11}\neq 0$ and $D_{22}\neq 0$ may be replaced with an equivalent problem in which $\hat{D}_{11}=0$ and $\hat{D}_{22}=0$. This observation is particularly useful in theoretical work requiring state-space calculations, since, as we will soon show, the case $D_{11}\neq 0$ leads to controllers with an unwieldy number of terms.

Step 1. The purpose of this step is to solve the four block problem at infinity. Suppose F_∞ is a constant feedback to be defined, and suppose also that $P(s)$ is as given in (5.1). Then for any such F_∞

(5.70)

$$\bar{P}(s)=F_l\left(\left[\begin{array}{cc|c} P_{11} & P_{12} & P_{12} \\ P_{21} & P_{22} & P_{22} \\ \hline P_{21} & P_{22} & P_{22} \end{array}\right], F_\infty\right)$$

$$=\left[\begin{array}{c|cc} \bar{A} & \bar{B}_1 & \bar{B}_2 \\ \hline \bar{C}_1 & \bar{D}_{11} & \bar{D}_{12} \\ \bar{C}_2 & \bar{D}_{21} & \bar{D}_{22} \end{array}\right]$$

$$=\left[\begin{array}{c|cc} A+B_2F_\infty(I-D_{22}F_\infty)^{-1}C_2 & B_1+B_2F_\infty(I-D_{22}F_\infty)^{-1}D_{21} & B_2(I-F_\infty D_{22})^{-1} \\ \hline C_1+D_{12}F_\infty(I-D_{22}F_\infty)^{-1}C_2 & D_{11}+D_{12}F_\infty(I-D_{22}F_\infty)^{-1}D_{21} & D_{12}(I-F_\infty D_{22})^{-1} \\ (I-D_{22}F_\infty)^{-1}C_2 & (I-D_{22}F_\infty)^{-1}D_{21} & (I-D_{22}F_\infty)^{-1}D_{22} \end{array}\right].$$

To find an F_∞ that solves the problem at infinity, we define $Q_\infty=F_\infty(I-D_{22}F_\infty)^{-1}$ and apply Parrott's result [39], which states that

$$Q_\infty=-D_{12}'(D_{11}+D_{11}\tilde{D}_\perp'(\gamma^2I-(D_\perp'D_{11}\tilde{D}_\perp')'(D_\perp'D_{11}\tilde{D}_\perp'))^{-1}(D_\perp'D_{11}\tilde{D}_\perp')'D_\perp'D_{11})D_{21}'$$

solves the problem

$$\left\|\begin{bmatrix} D_\perp'D_{11}\tilde{D}_\perp' & D_\perp'D_{11}D_{21}' \\ D_{12}'D_{11}\tilde{D}_\perp' & D_{12}'D_{11}D_{21}'+Q_\infty \end{bmatrix}\right\|_2<\gamma.$$

(A solution exists if and only if $\gamma>\gamma_p=\max\{\|D_\perp'D_{11}\|, \|D_{11}\tilde{D}_\perp'\|\}$). Back substitution gives

$$F_\infty=(I+Q_\infty D_{22})^{-1}Q_\infty$$

in which the existence of the inverse is assumed. There are two points to note:

(1) $\|\bar{D}_{11}\|_2<\gamma$.

(2) $I-D_{22}F_\infty=(I+D_{22}Q_\infty)^{-1}$ which shows that the existence of $(I+D_{22}Q_\infty)^{-1}\Rightarrow$ the existence of $(I-D_{22}F_\infty)^{-1}$.

Step 2. Here we select an orthogonal Θ-matrix in Fig. 1 such that $(F_l(\Theta, \bar{P}_{11}(s))(\infty))_{11}=0$. Note that

$$\bar{D}_{11} := D_{11} + D_{12}Q_\infty D_{21}$$

and define

$$\begin{bmatrix} \Theta_{11} & \Theta_{12} \\ \Theta_{21} & \Theta_{22} \end{bmatrix} = \gamma^{-1}\begin{bmatrix} \gamma^{-1}\bar{D}_{11} & (I-\gamma^{-2}\bar{D}_{11}\bar{D}'_{11})^{1/2} \\ -(I-\gamma^{-2}\bar{D}'_{11}\bar{D}_{11})^{1/2} & \gamma^{-1}\bar{D}'_{11} \end{bmatrix},$$

which satisfies $\Theta\Theta' = \gamma^{-2}I$ for all $\gamma \geqq \gamma_p$. By direct computation

(5.71)

$$\hat{P}(s) = F_l\left(\left[\begin{array}{cc|cc} \Theta_{11} & 0 & \Theta_{12} & 0 \\ 0 & 0 & 0 & I \\ \hline \Theta_{21} & 0 & \Theta_{22} & 0 \\ 0 & I & 0 & 0 \end{array}\right], \bar{P}(s)\right) \overset{s}{=} \left[\begin{array}{c|cc} \hat{A} & \hat{B}_1 & \hat{B}_2 \\ \hline \hat{C}_1 & 0 & \hat{D}_{12} \\ \hat{C}_2 & \hat{D}_{21} & \hat{D}_{22} \end{array}\right]$$

$$\overset{s}{=} \left[\begin{array}{c|cc} \bar{A}+\bar{B}_1\Theta_{22}(I-\bar{D}_{11}\Theta_{22})^{-1}\bar{C}_1 & \bar{B}_1(I-\Theta_{22}\bar{D}_{11})^{-1}\Theta_{21} & \bar{B}_2+\bar{B}_1\Theta_{22}(I-\bar{D}_{11}\Theta_{22})^{-1}\bar{D}_{12} \\ \hline \Theta_{12}(I-\bar{D}_{11}\Theta_{22})^{-1}\bar{C}_1 & 0 & \Theta_{12}(I-\bar{D}_{11}\Theta_{22})^{-1}\bar{D}_{12} \\ \bar{C}_2+\bar{D}_{21}\Theta_{22}(I-\bar{D}_{11}\Theta_{22})^{-1}\bar{C}_1 & \bar{D}_{21}(I-\Theta_{22}\bar{D}_{11})^{-1}\Theta_{21} & \bar{D}_{22}+\bar{D}_{21}\Theta_{22}(I-\bar{D}_{11}\Theta_{22})^{-1}\bar{D}_{12} \end{array}\right],$$

which has the required property that $\hat{D}_{11}=0$. It is an immediate consequence of a specialization of Redheffer's theorem that $\|F_l(\bar{P}, K)\|_\infty \leqq \gamma$ if and only if $\|F_l(\bar{P}, K)\|_\infty \leqq \gamma^{-1}$ [30]. A small gain argument shows that the internal stability property is preserved for all $\gamma > \gamma_p$.

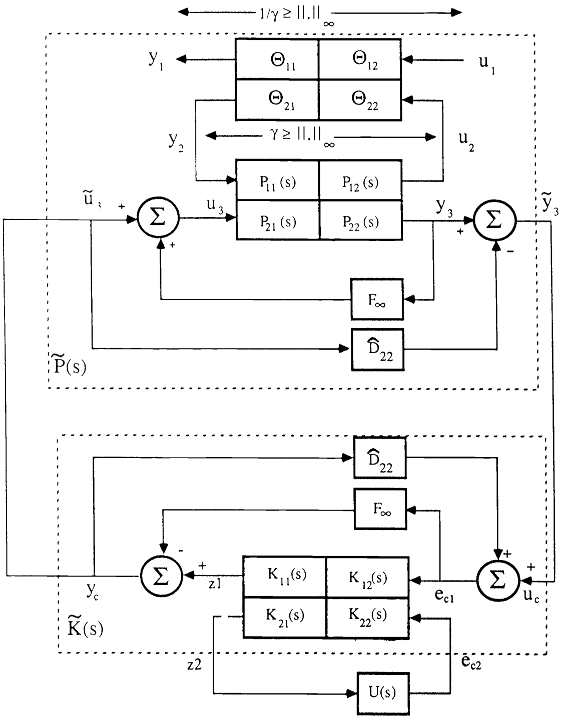


FIG. 1. *Loop transformations.*

Step 3. Here we eliminate $\hat{D}_{22}$ by connecting $(\hat{D}_{22}-\hat{D}_{22})$ in parallel with $\hat{P}_{22}(s)$ which is illustrated in Fig. 1. See also [20].

Step 4. Compute $\tilde{K}(s)$ using equations (5.11)-(5.18). (Note that $\hat{D}'_{12}\hat{D}_{12}\neq I$ and $\hat{D}_{21}\hat{D}'_{21}\neq I$ so another scaling is required.)

Step 5. Reverse the effects of $\hat{D}_{22}$ and F_∞ to obtain a representation formula for all controllers. Yet another calculation verifies that

$$\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} = F_l\left\{\left[\begin{array}{cc|c} F_\infty & \tilde{K}_{12} & -I \\ \tilde{K}_{21} & \tilde{K}_{22}-\tilde{K}_{21}\hat{D}_{22}\tilde{K}_{12} & \tilde{K}_{21}\hat{D}_{22} \\ \hline -I & \hat{D}_{22}\tilde{K}_{12} & -\hat{D}_{22} \end{array}\right], \tilde{K}_{11}\right\}.$$

Note that the effects of the scale introduced in Step 4 must now be reversed.

It remains to show that provided γ is big enough, the realization for $P(s)$ in (5.1) satisfies (i)-(iv) if and only if the realization for $\hat{P}(s)$ given in (5.70) satisfies these same assumptions. In our analysis below, we treat the transformation between $P(s)$ and $\bar{P}(s)$ first, and the transformation between $\bar{P}(s)$ and $\hat{P}(s)$ second.

LEMMA 5.4. *Assumptions* (i)-(iv) *apply to the realization in* (5.1) *if and only if they apply to the realization in* (5.70).

Proof. (i) Stabilizibility and detectability are invariant under output feedback. Assumption (ii) is immediate from (5.70). Assumption (iii) and (iv) follow from

$$\begin{bmatrix} sI-\bar{A} & -\bar{B}_2 \\ \bar{C}_1 & \bar{D}_{12} \end{bmatrix} = \begin{bmatrix} sI-A & -B_2 \\ C_1 & D_{12} \end{bmatrix}\begin{bmatrix} I & 0 \\ F_\infty(I-D_{22}F_\infty)^{-1}C_2 & (I-F_\infty D_{22})^{-1} \end{bmatrix}$$

$$\begin{bmatrix} sI-\bar{A} & \bar{B}_1 \\ -\bar{C}_2 & \bar{D}_{21} \end{bmatrix} = \begin{bmatrix} I & B_2F_\infty(I-D_{22}F_\infty)^{-1} \\ 0 & (I-D_{22}F_\infty)^{-1} \end{bmatrix}\begin{bmatrix} sI-A & B_1 \\ -C_2 & D_{21} \end{bmatrix}.$$

LEMMA 5.5. *Suppose that an internally stabilizing controller exists such that* $\|F_l(P,K)\|_\infty \leqq \gamma$. *Then*

(1) $(\hat{A}, \hat{B}_2, \hat{C}_2)$ *is stabilizable and detectable.*

(2) rank $(D_{12})=m \Leftrightarrow$ rank $(\hat{D}_{12})=m$, *and* rank $(D_{21}) \Leftrightarrow$ rank $(\hat{D}_{21})=q$.

(3) rank $\left(\left[\begin{smallmatrix} j\omega I-A & B_2 \\ -C_1 & D_{12} \end{smallmatrix}\right]\right)=n+m$ *for all real* ω *if and only if* rank $\left(\left[\begin{smallmatrix} j\omega I-\hat{A} & \hat{B}_2 \\ -\hat{C}_1 & \hat{D}_{12} \end{smallmatrix}\right]\right)=n+m$ *for all real* ω.

(4) rank $\left(\left[\begin{smallmatrix} j\omega I-A & -B_1 \\ C_2 & D_{21} \end{smallmatrix}\right]\right)=n+q$ *for all real* ω *if and only if* rank $\left(\left[\begin{smallmatrix} j\omega I-\hat{A} & -\hat{B}_1 \\ \hat{C}_2 & \hat{D}_{21} \end{smallmatrix}\right]\right)=n+q$ *for all real* ω.

Proof. (1) Since $\|\Theta_{22}\|_2<\gamma^{-1}$, it follows from a small gain argument that $F_l(\hat{P},K)$ is internally stable. As K is an internally stabilizing controller for $\hat{P}(s)$, it follows that $(\hat{A}, \hat{B}_2, \hat{C}_2)$ is stabilizable and detectable.

(2) This follows from the invertibility of Θ_{12} and Θ_{21}, (5.70) and (5.71).

Parts (3) and (4) follow from Lemma 5.3 and the identities

$$\begin{bmatrix} sI-\hat{A} & \hat{B}_2 \\ -\hat{C}_1 & \hat{D}_{12} \end{bmatrix} = \begin{bmatrix} I & \bar{B}_1\Theta_{22}(I-\bar{D}_{11}\Theta_{22})^{-1} \\ 0 & \Theta_{12}(I-\bar{D}_{11}\Theta_{22})^{-1} \end{bmatrix}\begin{bmatrix} sI-\bar{A} & \bar{B}_2 \\ -\bar{C}_1 & \bar{D}_{12} \end{bmatrix},$$

$$\begin{bmatrix} sI-\hat{A} & -\hat{B}_1 \\ \hat{C}_2 & \hat{D}_{21} \end{bmatrix} = \begin{bmatrix} sI-\bar{A} & -\bar{B}_1 \\ \bar{C}_2 & \bar{D}_{21} \end{bmatrix}\begin{bmatrix} I & 0 \\ \Theta_{22}(I-\bar{D}_{11}\Theta_{22})^{-1}\bar{C}_1 & (I-D_{22}\Theta_{22})^{-1}\Theta_{21} \end{bmatrix}. \quad \square$$

If we need to calculate a controller generator in terms of the original plant description, it is possible to repeat calculations of the type given in the previous two sections (these have only been checked in detail by the present authors in the case where X_∞ and Y_∞ exist). In this event Step 2 is left out and we proceed as follows: Execute Step 1 as before; note that this step can be carried out if and only if max $(\|D'_\perp D_{11}\|_2, \|D_{11}\tilde{D}'_\perp\|_2)<\gamma$. After this step there holds $\|D_{11}\|_2<\gamma$. Execute Step 3 as before to remove $\bar{D}_{22}$. Execute Step 4 using the results of Theorem 5.1′ and Theorem 5.2′ given below. End with Step 5 as before.

THEOREM 5.1′. *Suppose that $P(s)$ is given by* (5.1), *and that the assumptions* (i)–(iv) *are satisfied with $D_{22}=0$ and $\|D_{11}\|_2<\gamma$. Then internally stabilizing controllers exist for $\gamma>0$ with $\|F_l(P,K)\|_\infty \leqq \gamma$ if and only if*

(i) *Condition* (i) *of Theorem* 5.1 *is satisfied with* (5.5) *replaced by*

$$H_\infty=\begin{bmatrix} H_{\infty 11} & H_{\infty 12} \\ H_{\infty 21} & -H'_{\infty 11}\end{bmatrix}$$

in which

$$H_{\infty 11}=A+(B_1D'_{11}D_\perp D'_\perp-\gamma^2B_2D'_{12})(\gamma^2I-D_{11}D'_{11}D_\perp D'_\perp)^{-1}C_1$$
$$H_{\infty 21}=-C'_1D_\perp(I-\gamma^{-2}D'_\perp D_{11}D'_{11}D_\perp)^{-1}D'_\perp C_1$$
$$H_{\infty 12}=(B_1-B_2D'_{12}D_{11})(\gamma^2I-D'_{11}D_\perp D'_\perp D_{11})^{-1}(B_1-B_2D'_{12}D_{11})'-B_2B'_2$$

(ii) *Condition* (ii) *of Theorem* 5.1 *is satisfied with* (5.8) *replaced by*

$$J_\infty=\begin{bmatrix} J_{\infty 11} & J_{\infty 12} \\ J_{\infty 21} & -J'_{\infty 11}\end{bmatrix}$$

where

$$J_{\infty 11}=A+B_1(\gamma^2I-\tilde D'_\perp\tilde D_\perp D'_{11}D_{11})^{-1}(\tilde D'_\perp\tilde D_\perp D'_{11}C_1-\gamma^2D'_{21}C_2),$$
$$J_{\infty 21}=-B_1\tilde D'_\perp(I-\gamma^{-2}\tilde D_\perp D'_{11}D_{11}\tilde D'_\perp)^{-1}\tilde D_\perp B'_1,$$
$$J_{\infty 12}=(C_1-D_{11}D'_{21}C_2)'(\gamma^2I-D_{11}\tilde D'_\perp\tilde D_\perp D'_{11})^{-1}(C_1-D_{11}D'_{21}C_2)-C'_2C_2.$$

(iii) $$\begin{bmatrix} X'_{\infty 2}X_{\infty 1} & \gamma^{-1}X'_{\infty 2}Y_{\infty 2} \\ \gamma^{-1}Y'_{\infty 2}X_{\infty 2} & Y'_{\infty 2}Y_{\infty 1}\end{bmatrix}\geqq 0.$$ □

All the solutions may be characterized by Theorem 5.2′.

THEOREM 5.2′. *If the conditions of Theorem* 5.1′ *are satisfied, then all internally stabilizing controllers satisfying $\|F_l(P,K)\|_\infty\leqq\gamma$ are given by*

$$K=F_l(\mathscr{K},U)\quad \textit{with } U\in RH^\infty_+,\quad \|U\|_\infty\leqq\gamma$$

where

$$\mathscr{K}(s)=\begin{bmatrix} 0 & D_{k12} \\ D_{k21} & D_{k22}\end{bmatrix}+\begin{bmatrix} C_{k1} \\ C_{k2}\end{bmatrix}(sE_k-A_k)^{\#}[B_{k1}\quad B_{k2}],$$

$$E_k=Y'_{\infty 1}X_{\infty 1}-\gamma^{-2}Y'_{\infty 2}X_{\infty 2},$$
$$B_{k1}=(\gamma^2Y'_{\infty 1}B_1+Y'_{\infty 2}C'_1D_{11}+Y'_{\infty 2}C'_2D_{21}(\gamma^2I-D'_{11}D_{11}))(\gamma^2I-\tilde D'_\perp\tilde D_\perp D'_{11}D_{11})^{-1}D'_{21},$$
$$C_{k1}=-D'_{12}(\gamma^2I-D_{11}D'_{11}D_\perp D'_\perp)^{-1}(\gamma^2C_1X_{\infty 1}+D_{11}B'_1X_{\infty 2}$$
$$+(\gamma^2I-D_{11}D'_{11})D_{12}B'_2X_{\infty 2}),$$
$$D_{k12}=\{I-D'_{12}D_{11}(\gamma^2I-D'_{11}D_\perp D'_\perp D_{11})^{-1}D'_{11}D_{12}\}^{1/2},$$
$$D_{k21}=\{I-D_{21}D'_{11}(\gamma^2I-D_{11}\tilde D'_\perp\tilde D_\perp D'_{11})^{-1}D_{11}D'_{21}\}^{1/2},$$
$$D_{k22}=-(D^{-1}_{k21})'D_{21}D'_{11}(\gamma^2I-D_{11}\tilde D'_\perp\tilde D_\perp D'_{11})^{-1}D_{12}D_{k12},$$
$$A_k=E_kT_x+B_{k1}D^{-1}_{k21}C_{k2}$$
$$=T'_YE_k+B_{k2}D^{-1}_{k12}C_{k1},$$
$$B_{k2}=\{Y'_{\infty 1}B_2+(Y'_{\infty 1}B_1\tilde D'_\perp\tilde D_\perp D'_{11}+Y'_{\infty 2}(C'_1-C'_2D_{21}D'_{11}))$$
$$\cdot(\gamma^2I-D_{11}\tilde D'_\perp\tilde D_\perp D'_{11})^{-1}D_{12}\}D_{k12},$$
$$C_{k2}=-D_{k21}\{C_2X_{\infty 1}+D_{21}(\gamma^2I-D'_{11}D_\perp D'_\perp D_{11})^{-1}(D'_{11}D_\perp D'_\perp C_1X_{\infty 1}+B'_1X_{\infty 2}$$
$$-D'_{11}D_{12}B'_2X_{\infty 2})\}.$$ □

6. Conclusions. The purpose of this paper was to derive a representation formula for all H_∞^+ solutions to the four block general distance problem in which particular attention was paid to the optimal cases. The suboptimal case was treated in § 3, and was derived by an all-pass embedding procedure which is reminiscent of that introduced by Parrott [39]. The optimal case was treated in § 4 and the formula for all solutions appears as (4.52) and (4.59). Section 5 dealt with the application to H^∞ optimal control. In addition, we note the following:

(1) In contrast to the one block problem, there are two types of optimality. The first is essentially the same as that addressed in Parrott's theorem [39], and has nothing to do with the requirement that the approximation Q be an element of H_∞^+. The second is associated with a Hankel norm condition on R_a in (3.46). In the first type of optimality, special algorithms may be required to take care of axis phenomena in the spectral factorization problems (3.18) and (3.19). These issues are addressed in detail in [8].

(2) The analysis in § 3 breaks down when $\|R_a\|_H = 1$ since Z in (3.65) becomes singular. This difficulty is addressed in § 4 where we introduce an alternative construction based on descriptor representations of all-pass transfer functions; see also [35] and [44] in this connection. In the case of this Hankel norm type of optimality, the effective dimension of the free parameter drops by rank $(B_{22}) \leqq k$. B_{22} is defined in (4.38) and k is the multiplicity of the largest Hankel singular value of R_a. This loss of effective dimension may be characterized by the linear constraining equation (4.45) which is similar to that given in [16, eq. (6.23)].

(3) An important application of this work is the derivation of a closed form representation formula for all controllers that satisfy an L^∞ norm constraint. One such formula is given in § 5 and appears in equations (5.10)-(5.18) in the case that $D_{11} = 0$ and $D_{22} = 0$ in the realization of $P(s)$ given in (5.1). All the optimal cases are covered.

(4) The case in which $D_{11} \neq 0$ has been treated by direct calculation and the results are given in Theorem 5.2′. Since these formulae are awkward to write down, an alternative approach based on loop shifting is also presented [45]. The idea here is to replace the original problem which has $D_{11} \neq 0$ with an equivalent problem in which $\hat{D}_{11} = 0$.

REFERENCES

[1] B. D. O. Anderson, *An algebraic solution to the spectral factorization problem*, IEEE Trans. Automat. Control, 12 (1967), pp. 410-414.

[2] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis a Modern Systems Approach*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[3] J. A. Ball and N. Cohen, *Sensitivity minimization in an H^∞ norm: parameterization of all suboptimal solutions*, Internat. J. Control, 46 (1987), pp. 785-816.

[4] M. Banker, *Linear stationary quadratic games*, in Proc. Conference on Decision and Control, 1973, pp. 193-197.

[5] D. S. Bernstein and W. M. Haddad, LQG *control with an H^∞ performance bound: a Riccati equation approach*, IEEE Trans. Automat. Control, 34 (1989), pp. 293-305.

[6] M. Bettayeb, L. M. Silverman, and M. G. Safonov, *Optimal approximation of continuous time systems*, in Proc. IEEE Conference on Decision and Control, Albuquerque, NM, December 10-12, 1980.

[7] C. C. Chu, J. C. Doyle, and E. B. Lee, *The general distance problem in H^∞ optimal control theory*, Internat. J. Control, 44 (1986), pp. 565-596.

[8] D. J. Clements and K. Glover, *Spectral factorization via Hermitian pencils*, Linear Algebra Appl., Vol. 122/123/124 (1989), pp. 797-846.

[9] C. DAVIS, W. M. KAHAN, AND H. F. WEINBERGER, *Norm-preserving dilations and their applications to optimal error bounds*, SIAM J. Control Optim., (1982), pp. 445-469.

[10] J. C. DOYLE et al, *Lecture notes in advances in multivariable control*, ONR/Honeywell Workshop, Minneapolis, MN, 1984.

[11] J. C. DOYLE, K. GLOVER, P. KHARGONEKAR, AND B. FRANCIS, *State-space solutions to standard H_2 and H_∞ control problems* IEEE American Control Conference, Atlanta, GA, June 1988; for an extended version see IEEE Trans. Automat. Control, 34 (1989), pp. 831-847.

[12] H. DYM AND I. GOHBERG, *A new class of contractive interpolants and maximum entropy principles*, Birkhauser-Verlag, Berlin, 1988, pp. 117-150.

[13a] C. FOIAS AND A. TANNENBAUM, *On the four block problem* I, Operator Theory Adv. and Appl., 32 (1988), pp. 93-122.

[13b] ———, *On the four block problem* II: *The singular system*, Integral Equations and Operator Theory, 11 (1988), pp. 726-767.

[14] B. A. FRANCIS, *A course in H^∞ control theory*, Lecture Notes in Control and Information Sciences, Vol. 88, Springer-Verlag, Berlin, New York, 1987.

[15] B. A. FRANCIS AND J. C. DOYLE, *Linear control theory with an H^∞ optimality criterion*, SIAM J. Control Optim., 25 (1987), pp. 815-844.

[16] K. GLOVER, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds*, Internat. J. Control, 39 (1984), pp. 1115-1193.

[17] ———, *Model reduction*: *A tutorial on Hankel-norm methods and lower bounds on L^2 errors*, IFAC Congress, Munich, July, 1987.

[18] ———, *A tutorial on Hankel norm approximation*, in From Data to Model, J. C. Willems, ed., Springer-Verlag, Berlin, New York, 1989, pp. 26-48.

[19] K. GLOVER, R. F. CURTAIN, AND J. R. PARTINGTON, *Realisation and approximation of linear infinite dimensional systems with error bounds*, SIAM J. Control Optim., 26 (1988), pp. 863-898.

[20] K. GLOVER AND J. C. DOYLE, *State-space formulae for all stabilizing controllers that satisfy a H^∞ norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167-172.

[21] K. GLOVER AND D. MUSTAFA, *Derivation of the maximum entropy H^∞ controller and a state-space formula for its entropy*, Internat. J. Control, 50 (1989), pp. 899-916.

[22] M. GREEN, K. GLOVER, D. J. N. LIMEBEER, AND J. C. DOYLE, *A j-spectral factorization approach to H_∞ control*, SIAM J. Control Optim., 28 (1990), pp. 1350-1371.

[23] Y. S. HUNG, *H^∞ optimal control: part* I *model matching*, Internat. J. Control, 49 (1989), pp. 1331-1359.

[24] ———, *H^∞ optimal control—Part* II *solutions for controllers*, Internat. J. Control, 49 (1989), pp. 1331-1359.

[25] D. H. JACOBSON, *Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games*, IEEE Trans. Automat. Control, 18 (1973), pp. 124-131.

[26] E. A. JONCKHEERE AND J. C. JUANG, *Fast computation of achievable performance in mixed sensitivity H^∞ design*, IEEE Trans. Automat. Control., 32 (1987), pp. 896-906.

[27] E. M. KASENALLY AND D. J. N. LIMEBEER, *Closed formulae for a parametric mixed sensitivity problem*, Systems Control Lett., 12 (1989), pp. 1-7.

[28] P. P. KHARGONEKAR, I. R. PETERSEN, AND M. A. ROTEA, *H^∞ optimal control with state feedback*, IEEE Trans. Automat. Control, 33 (1988), pp. 783-786.

[29] H. KIMURA AND R. KAWATANI, *Synthesis of H^∞ controllers based on conjugation*, Proc. IEEE Conference on Decision and Control, Austin, TX, 1988.

[30] D. J. N. LIMEBEER AND B. D. O. ANDERSON, *An interpolation theory approach to H^∞ controller degree bounds*, Linear Algebra Appl., 98 (1988), pp. 347-386.

[31] D. J. N. LIMEBEER AND G. D. HALIKIAS, *An analysis of pole zero cancellations in H^∞-optimal control problems of the second kind*, SIAM J. Control Optim., 26 (1988), pp. 646-677.

[32] D. J. N. LIMEBEER AND Y. S. HUNG, *An analysis of pole-zero cancellations in H^∞-optimal control problems of the first kind*, SIAM. J. Control Optim., 25 (1987), pp. 1457-1493.

[33] D. J. N. LIMEBEER AND E. M. KASENALLY, *Vector interpolation*: *Applications to H^∞*, IMA Conference on Matrix Theory, Bradford, July 1988.

[34] D. J. N. LIMEBEER, B. D. O. ANDERSON, P. P. KHARGONEKAR, AND M. GREEN, *A game theoretic approach to H^∞ control for time varying systems*, SIAM J. Control Optim., submitted.

[35] D. J. N. LIMEBEER, E. M. KASENALLY, E. JAIMOUKA, AND M. G. SAFONOV, *A characterization of all solutions to the four block general distance problem*, in Proc. IEEE Conference on Decision and Control, Austin, TX, 1988.

[36] D. J. N. LIMEBEER, G. D. HALIKIAS, AND K. GLOVER, *A state-space algorithm for the computation of superoptimal matrix interpolating functions*, Linear Circuits, Systems and Signal Processing: Theory and Applications, North-Holland, Amsterdam, 1988.

[37] D. G. LUENBERGER, *Dynamic equations in descriptor form*, IEEE Trans. Automat. Control, 22 (1977), pp. 312-321.

[38] D. MUSTAFA, K. GLOVER, AND D. J. N. LIMEBEER, *Controllers which satisfy a closed-loop H^∞-norm bound and maximize an entropy integral*, Automatica, to appear.

[39] S. PARROTT, *On a quotient norm and the Sz.-Nagy-Foias lifting theorem*, J. Func. Anal., 30 (1978), pp. 311-328.

[40] I. R. PETERSEN AND D. J. CLEMENTS, *J-spectral factorization and Riccati equations in problems of H^∞ optimization via state feedback*, preprint, 1988.

[41] S. C. POWER, *Hankel Operators on Hilbert Space*, Pitman Advanced Publishing Program, Boston, 1981.

[42] R. M. REDHEFFER, *On a certain linear fractional transformation*, J. Math. Phys., 39 (1960), pp. 269-286.

[43] M. G. SAFONOV, E. A. JONCKHEERE, M. VERMA, AND D. J. N. LIMEBEER, *Systhesis of positive real multivariable feedback systems*, Internat. J. Control, 45 (1987), pp. 817-842.

[44] M. G. SAFONOV, R. Y. CHIANG, AND D. J. N. LIMEBEER, *Hankel model reduction without balancing: A descriptor system approach*, IEEE Conference on Decision and Control, Los Angeles, CA, 1987; an extended version to appear in IEEE Trans. Automat. Control.

[45] M. G. SAFONOV, D. J. N. LIMEBEER, AND R. Y. CHIANG *Simplifying the H^∞ theory via loop shifting, matrix pencil and descriptor concepts*, Internat. J. Control., 50 (1990), pp. 2467-2488.

[46] M. G. SAFONOV AND M. VERMA, *Multivariable L^∞ sensitivity optimization and Hankel approximation*, in Proc. American Control Conference, San Francisco, CA, 1983; also in IEEE Trans. Automat. Control, 30 (1985), pp. 279-280.

[47] G. TADMOR, *Worst-case design in the time domain; the maximum principle and the standard H_∞ problem*, Math. of Control Signals and Systems, to appear.

[48] M. S. VERMA AND J. C. ROMIG, *Reduced order controllers in H^∞-optimal synthesis methods of the first kind*, December, 1987, submitted.

[49] P. WHITTLE, *Risk sensitive linear quadratic gaussian control*, Adv. Appl. Probab., 13 (1981), pp. 764-777.

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

[51] H. WIMMER, *Monotonicity of maximal solutions of algebraic Riccati equations*, Systems Control Lett., 5 (1985), pp. 317-319.

[52] W. M. WONHAM, *Linear Multivariable Control: A Geometric Approach*, Series on Applications of Mathematics, Springer-Verlag, Berlin, New York, 1979.

[53] N. YOUNG, *An Introduction to Hilbert Space*, Cambridge University Press, London, 1988.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 325-338, March 1991

004

DUALITY RELATIONSHIPS FOR ENTROPY-LIKE MINIMIZATION PROBLEMS*

J. M. BORWEIN† AND A. S. LEWIS‡

Abstract. This paper considers the minimization of a convex integral functional over the positive cone of an L_p space, subject to a finite number of linear equality constraints. Such problems arise in spectral estimation, where the objective function is often entropy-like, and in constrained approximation. The Lagrangian dual problem is finite-dimensional and unconstrained. Under a quasi-interior constraint qualification, the primal and dual values are equal, with dual attainment. Examples show the primal value may not be attained. Conditions are given that ensure that the primal optimal solution can be calculated directly from a dual optimum. These conditions are satisfied in many examples.

Key words. convex programming, duality, spectral estimation, entropy, moment problem

AMS(MOS) subject classifications. primary 90C25, 49A27; secondary 41A46, 42A70

1. Introduction. A number of authors have recently considered dual approaches to the solution of optimization problems of the form

$$\text{(EP)}\qquad \begin{array}{ll} \inf & \displaystyle\int_T \phi(x(t))\,dt \\ \text{subject to} & Ax=b, \\ & x\geqq 0, \qquad x\in L_p(T). \end{array}$$

Here, (T, dt) is a σ-finite measure space, $1\leqq p\leqq\infty$, $\phi:\mathbb{R}\to(-\infty,\infty]$ is convex, and $A: L_p\to\mathbb{R}^n$ is continuous. Such problems arise in a number of areas. When (for $p<\infty$), $\phi(x)=(1/p)x^p$ we obtain the constrained L_p approximation problem (see [Micchelli, Smith, Swetits, and Ward, 1985]). These problems appear in the theory of constrained interpolation [Irvine, Marin, and Smith, 1986], and also as a result of using the L_p norm as the objective function in spectral estimation problems (see, for example, [Goodrich and Steinhardt, 1986] and [Ben-Tal, Borwein, and Teboulle, 1988a]). A number of different objective functions can be used. Typically, they are "entropic" in form, for instance $\phi(x)=-\log x$ [Burg, 1975] and $\phi(x)=x\log x$ [Johnson and Shore, 1984]. A survey of common objective functions may be found in [Ben-Tal, Borwein, and Teboulle, 1988b]. For simplicity of exposition we will only consider the autonomous case, where the function ϕ does not depend explicitly on the variable t. The non-autonomous case is a simple extension.

Let us denote the function $\phi+\delta(\cdot|\mathbb{R}_+)$ by ϕ_+ (where δ denotes an indicator function [Rockafellar, 1970]). Assuming ϕ_+ is a *normal convex integrand* in the sense of [Rockafellar, 1968], we can write the Lagrangian dual problem as

$$\text{(DEP)}\qquad \begin{array}{ll} \sup & \displaystyle b^T\lambda-\int_T (\phi_+)^*((A^T\lambda)(t))\,dt \\ \text{subject to} & \lambda\in\mathbb{R}^n, \end{array}$$

* Received by the editors May 31, 1988; accepted for publication (in revised form) January 22, 1990. This research was supported by the Natural Sciences and Engineering Research Council of Canada.

† Department of Mathematics, Statistics, and Computing Science, Dalhousie University, Halifax, Nova Scotia, Canada B3H 3J5.

‡ Department of Combinatorics and Optimization, Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1.

where * denotes convex conjugation and $A^T:\mathbb{R}^n \to L_q(T)$ $((1/p)+(1/q)=1)$ is the adjoint map. It is known (see, for example, [Ben-Tal, Borwein, and Teboulle, 1988b]) that under a suitable constraint qualification the values of (EP) and (DEP) are equal, with dual attainment.

Suppose now that ϕ_+ is closed and essentially strictly convex, so, by [Rockafellar, Thm. 26.3, 1970], $(\phi_+)^*$ is essentially smooth. (We defer precise definitions to a later section.) Suppose that $\bar{\lambda}$ is optimal for (DEP). *Assuming that* $(A^T\bar{\lambda})(t) \in \operatorname{int}(\operatorname{dom}(\phi_+)^*)$, *and that we can differentiate through the integral*, we then obtain

$$b - A(((\phi_+)^*)'((A^T\bar{\lambda})(\cdot))) = 0. \tag{1.1}$$

If we now set $\bar{x}(t) = ((\phi_+)^*)'((A^T\bar{\lambda})(t))$, $t \in T$, then by [Rockafellar, Thm. 23.5, 1970],

$$(A^T\bar{\lambda})(t) \in \partial\phi_+(\bar{x}(t)), \quad \text{a.e.},$$

so $\phi_+(\bar{x}(t)) + (\phi_+)^*((A^T\bar{\lambda})(t)) = \bar{x}(t)(A^T\bar{\lambda})(t)$, almost everywhere and, since $(\phi_+)^*$ is increasing, $\bar{x} \geqq 0$. Thus, by (1.1) $\bar{x}(t)$ is feasible for (EP) and, integrating over T, has the same objective value as the dual and hence is optimal (by weak duality).

The question of when the above assumptions are justified has, to the authors' knowledge, never been addressed in the published literature on the subject. Since the derivation of primal solutions is of paramount importance, this question is clearly extremely significant. The aim of this paper is therefore to give these matters a rigorous treatment. We begin by deriving the basic duality result from the theory developed in [Borwein and Lewis, 1988]. We then give some examples to show how the above assumptions can fail. This will motivate our rigorous treatment of the derivation of primal solutions.

2. Duality.

We will begin by deriving the fundamental duality result.

DEFINITION 2.1 [Borwein and Lewis, 1988]. Let (X, τ) be a topological vector space, with convex $C \subset X$. The *quasi-relative interior* of C $(\tau - \operatorname{qri} C)$ is the set of those $x \in C$ for which $\operatorname{cl} \mathbb{P}(C - x)$ is a subspace.

(Here, $\operatorname{cl} \mathbb{P}B$ denotes the closed cone generated by B.) Note that if X is normed, the weak and norm quasi-relative interiors coincide.

Example 2.2 [Borwein and Lewis, 1988]. Let T be a σ-finite measure space, $1 \leqq p \leqq \infty$, $(1/p)+(1/q)=1$ and let $X = L_p(T)$, $Y = L_q(T)$. Then

$$\sigma(X, Y) - \operatorname{qri} X_+ = \{x \mid x(t) > 0 \text{ a.e.}\}.$$

LEMMA 2.3. *Let X be a topological vector space with convex C_1, $C_2 \subset X$ satisfying $\operatorname{cl} C_1 = \operatorname{cl} C_2$. For any $x \in C_1 \cap C_2$, $x \in \operatorname{qri} C_1$ if and only if $x \in \operatorname{qri} C_2$.*

Proof. The result follows immediately from the fact that $\operatorname{cl} \mathbb{P}(C_1 - x) = \operatorname{cl} \mathbb{P}(C_2 - x)$. □

The following duality result may be found in [Borwein and Lewis, Cor. 4.10, 1988].

THEOREM 2.4. *Let X be locally convex, $f: X \to (-\infty, \infty]$ convex, $A: X \to \mathbb{R}^n$ continuous and linear, $b \in \mathbb{R}^n$, and $P \subset \mathbb{R}^n$ a polyhedral cone. Consider the following dual pair of problems*:

$$\text{(CM)} \qquad \begin{array}{ll} \inf & f(x) \\ \textit{subject to} & Ax \in b + P, \\ & x \in C, \end{array}$$

$$\text{(DCM)} \qquad \begin{array}{ll} \max & b^T\lambda - (f + \delta(\cdot \mid C))^*(A^T\lambda) \\ \textit{subject to} & \lambda \in P^+. \end{array}$$

If there exists a feasible $\hat{x} \in \operatorname{qri}((\operatorname{dom} f) \cap C)$ *for* (CM), *then the values of* (CM) *and* (DCM) *are equal* (*with attainment in* (DCM)).

To use the above result in our case we need to compute the conjugate of certain integral functionals. We will need the following result concerning normal convex integrands.

THEOREM 2.5. *Let T be a finite measure space, $1 \leqq p \leqq \infty$, and suppose $\psi : \mathbb{R} \to (-\infty, \infty]$ is closed, convex, and proper. Define $I_\psi : L_p(T) \to (-\infty, \infty]$ by $I_\psi(x) = \int_T \psi(x(t))\, dt$. Then $(I_\psi)^* : L_q(T) \to (-\infty, \infty]$ is given by $(I_\psi)^* = \int_T \psi^*(y(t))\, dt$.*

Proof. For the proof see the corollary to Theorem 2 in [Rockafellar, 1968]. □

We can now derive the required duality result.

COROLLARY 2.6. *Suppose T is a finite measure space, $1 \leqq p \leqq \infty$, $\phi : \mathbb{R} \to (-\infty, \infty]$ with ϕ_+ closed and convex, $a_i \in L_q(T)$ for $i = 1, \cdots, n$, and $b \in \mathbb{R}^n$. Consider the following dual pair of problems*:

$$(\mathrm{EP}_p) \qquad \begin{aligned} &\inf && \int_T \phi(x(t))\, dt \\ &\textit{subject to} && \int_T x(t) a_i(t)\, dt = b_i, \qquad i = 1, \cdots, n, \\ & && x \geqq 0, \qquad x \in L_p(T), \end{aligned}$$

$$(\mathrm{DEP}_p) \qquad \begin{aligned} &\max && b^T\lambda - \int_T (\phi_+)^* \left[\sum_{i=1}^n \lambda_i a_i(t) \right] dt \\ &\textit{subject to} && \lambda \in \mathbb{R}^n. \end{aligned}$$

Suppose further that the following constraint qualification holds:

(CQ) *There exists a feasible $\hat{x}$ for* (EP_p) *with $\hat{x} \in \sigma(L_p, L_q) - \operatorname{qri}(\operatorname{dom} I_\phi \cap (L_p)_+)$.*

Then the values of (EP_p) *and* (DEP_p) *are equal* (*with dual attainment*).

Suppose furthermore that $(0, \infty) \subset \operatorname{dom} \phi$. Then (CQ) *is equivalent to requiring the existence of a feasible $\hat{x} \in \operatorname{dom} I_\phi$ with $\hat{x}(t) > 0$ almost everywhere. In particular, if* $\operatorname{ess\,sup} x < +\infty$ *and* $\operatorname{ess\,inf} x > 0$ *then* $x \in \operatorname{dom} I_\phi$.

Proof. In Theorem 2.4, set $X = L_p(T)$ with the $\sigma(L_p, L_q)$ topology, $f = I_\phi$, $C = (L_p)_+$, A defined by $(Ax)_i = \langle x, a_i \rangle = \int_T x(t) a_i(t)\, dt$ for $i = 1, \cdots, n$, and $P = \{0\}$. Then $f + \delta(\cdot \mid C) = I_{\phi_+}$, so by Theorem 2.5, $(f + \delta(\cdot \mid C))^* = I_{\phi_+^*}$. Also, $P^+ = \mathbb{R}^n$, and $A^T\lambda = \sum_{i=1}^n \lambda_i a_i$, so the duality result follows.

If $(0, \infty) \subset \operatorname{dom} \phi$ then the set $K := \{x \in L_\infty \mid \operatorname{ess\,inf} x > 0\}$ is contained in $\operatorname{dom} I_\phi \cap (L_p)_+$. To see this, observe that if $x \in L_\infty$ with $\operatorname{ess\,inf} x > 0$, then there exist $M, \varepsilon > 0$ with $\varepsilon \leqq x(t) \leqq M$, almost everywhere. Since $(0, \infty) \subset \operatorname{dom} \phi$, ϕ is continuous on $(0, \infty)$, by [Rockafellar, Thm. 10.1, 1970], and so the set $\phi[\varepsilon, M] \subset \mathbb{R}$ is compact. Thus, $\phi(x(\cdot))$ is bounded on T, and so $x \in \operatorname{dom} I_\phi$.

Now K is dense in $(L_p)_+$ (the proof essentially follows [Rudin, Thm. 3.13, 1966]). Thus, we have

$$\begin{aligned} \operatorname{cl} K &\subset \operatorname{cl}(\operatorname{dom} I_\phi \cap (L_p)_+) \\ &\subset \operatorname{cl}((L_p)_+) \\ &= (L_p)_+ \\ &= \operatorname{cl} K, \end{aligned}$$

so $\operatorname{cl}(\operatorname{dom} I_\phi \cap (L_p)_+) = (L_p)_+$. Applying Lemma 2.3 it follows that if $\hat{x} \in (\operatorname{dom} I_\phi) \cap \operatorname{qri}(L_p)_+$ then $\hat{x} \in \operatorname{qri}((\operatorname{dom} I_\phi) \cap (L_p)_+)$, and the result follows by Example 2.2. □

In the case where the measure space is nonatomic and totally finite, we can handle nonconvex objectives (as observed by the referee). Suppose that ϕ_+ is lower semicontinuous (not necessarily convex), but that ϕ_+^* and ϕ_+^{**} are proper. Then, with A as before, if we define

$$h(b) := \inf\left\{\int_T \phi_+(x(t))\,dt \,|\, Ax=b,\ x\in L_1\right\},$$

the domain of h is convex, and h is a convex function. This is a consequence of the Liapunov convexity theorem [Holmes, p. 108, 1975]. In fact if we define the value function of the regularized problem,

$$\bar{h}(b) := \inf\left\{\int_T \phi_+^{**}(x(t))\,dt \,|\, Ax=b,\ x\in L_1\right\},$$

then whenever the constraint qualification is satisfied at b, in other words $b\in A(\mathrm{qri}\,(\mathrm{dom}\, I_{\phi_+}))$, we have $h(b)=\bar{h}(b)$. This follows from [Rockafellar, Thm. 3H, 1976], and the fact that the constraint qualification forces h to be lower semicontinuous at b. In conclusion, our results apply equally well to ϕ as to ϕ^{**}. See also [Ioffe and Tihomorov, 1968] for similar ideas.

In some common cases the constraint qualification is particularly easy to check. The next result gives conditions under which the constraint qualification is no more arduous than primal consistency. Note that in most applications the a_i's are actually continuous (at least piecewise).

DEFINITION 2.7 [Borwein and Lewis, 1988]. A set of measurable functions $a_i: T\to\mathbb{R}$, $i=1,\cdots,n$, on a measure space T are called *pseudo-Haar* if they are linearly independent on every nonnull subset of T.

PROPOSITION 2.8. *Consider the measure space* $[\alpha,\beta]\subset\mathbb{R}$, *with Lebesgue measure, and suppose* $a_i:[\alpha,\beta]\to\mathbb{R}$ *are analytic and linearly independent. Then they are pseudo-Haar on* $[\alpha,\beta]$.

(A real function f defined on an open interval in $\mathbb{R}$ is called *analytic* if it is represented locally by an absolutely convergent power series at any point in the interval, cf. [Bochner and Martin, 1964].)

Proof. We will show that if a real function f is analytic on an open interval $T\supset[\alpha,\beta]$, and is not identically zero on $[\alpha,\beta]$, then it has at most finitely many zeros on $[\alpha,\beta]$. The result then follows immediately.

Suppose f has infinitely many zeros on $[\alpha,\beta]$. By compactness, there exists a sequence $x_i\to x_0\in[\alpha,\beta]$ as $i\to\infty$, with $f(x_i)=0$, $x_i\neq x_0$, for $i=1,2,\cdots$. Since f is analytic, for some $\delta>0$ there is an absolutely convergent power series $f(x)=\sum_{n=0}^{\infty} c_n(x-x_0)^n$, for $x\in N_\delta(x_0)$. Suppose the c_n's are not all 0, and let c_m be the first nonzero coefficient. Then

$$f(x)=(x-x_0)^m\left\{c_m+(x-x_0)\sum_{n=m+1}^{\infty} c_n(x-x_0)^{n-m-1}\right\},$$

for $x\in N_\delta(x_0)$. Since the series $\sum_{n=m+1}^{\infty} c_n(x-x_0)^{n-m-1}$ converges absolutely on $N_\delta(x_0)$, by a standard power series result it converges uniformly on $N_{\delta/2}(x_0)$, so in particular it is continuous at x_0. We then have

$$\begin{aligned}0&=(x_i-x_0)^{-m}f(x_i)\\ &=c_m+(x_i-x_0)\sum_{n=m+1}^{\infty} c_n(x_i-x_0)^{n-m-1},\end{aligned}$$

and letting $i\to\infty$ implies $c_m=0$, which is a contradiction.

Thus, $c_n = 0$ for all n, so $f(x) = 0$ on $N_\delta(x_0)$. But 0 is an analytic function on the domain T which agrees with f on $N_\delta(x_0)$, so by [Bochner and Martin, Thm. 4, Chap. II, 1964], $f \equiv 0$ on T. This is a contradiction. □

The following result shows essentially that if the constraint functions a_i are pseudo-Haar and (EP_p) is consistent for any $1 \leqq p \leqq \infty$ then there is in fact a feasible $\hat{x}$ in the norm interior of $(L_\infty)_+$. In the case when $(0, \infty) \subset \operatorname{dom} \phi$, $\hat{x}$ clearly satisfies the constraint qualification, so the result shows that the constraint qualification is always satisfied for a consistent primal (providing 0 is not the only primal feasible solution).

THEOREM 2.9. *Suppose (T, μ) is a finite measure space and $0 \leqq x \in L_p(T)$ is nonzero $(1 \leqq p \leqq \infty)$. Suppose further that $a_i \in L_q(T)$, $i = 1, \cdots, n$, are pseudo-Haar. Then there exists a $y \in L_\infty(T)$ and $\varepsilon > 0$ with $y(t) \geqq \varepsilon$ almost everywhere and $\langle x, a_i\rangle = \langle y, a_i\rangle$, each i.*

Proof. Since $0 \neq x \geqq 0$, there exists $T_1 \subset T$ with $\mu(T_1) > 0$ and a $\delta > 0$ such that $x(t) \geqq \delta$ almost everywhere on T_1. We claim that

$$\left\{\left(\int_{T_1} u a_i \, d\mu\right)_{i=1}^n \,\middle|\, u \in L_\infty(T_1)\right\} = \mathbb{R}^n. \tag{2.10}$$

Suppose not. Since the left-hand side is clearly a subspace, there exists $\lambda \neq 0$ such that $\sum_{i=1}^n \lambda_i \int_{T_1} u a_i \, d\mu = 0$, for all $u \in L_\infty(T_1)$. This implies $\sum_{i=1}^n \lambda_i a_i(t) = 0$ almost everywhere on T_1, which contradicts the fact that the a_i's are pseudo-Haar. Thus, (2.10) holds.

Now define

$$C = \left\{\left(\int_{T_1} u a_i \, d\mu\right)_{i=1}^n \,\middle|\, u \in L_\infty(T_1),\ \|u\|_\infty < \delta/2\right\}.$$

Since, by (2.10), $\mathbb{P}C = \mathbb{R}^n$, it follows that $0 \in \operatorname{int} C$ [Rockafellar, Cor. 6.4.1, 1970].

Let us now define a sequence of functions $(x_m) \subset L_\infty(T)$ by

$$x_m(t) = \begin{cases} m, & \text{if } x(t) > m, \\ x(t), & \text{if } \dfrac{1}{m} \leqq x(t) \leqq m, \\ \dfrac{1}{m}, & \text{if } x(t) < \dfrac{1}{m}. \end{cases}$$

For $p < \infty$ we have

$$\|x_m - x\|_p^p \leqq \int_{\{t \mid x(t) > m\}} x(t)^p \, d\mu + \int_{\{t \mid x(t) < 1/m\}} \left(\frac{1}{m}\right)^p d\mu \to 0, \quad \text{as } m \to \infty,$$

and in the case $p = \infty$, clearly $\|x_m - x\|_\infty \leqq 1/m$, for m large. Thus, $((\langle x_m - x, a_i\rangle))_{i=1}^n \to 0$ as $m \to \infty$, and so, since $0 \in \operatorname{int} C$, for $m > 1/\delta$ sufficiently large, $((x_m - x, a_i\rangle)_{i=1}^n \in C$. It then follows that $\langle x_m - x, a_i\rangle = \langle v, a_i\rangle$, each $i = 1, \cdots, n$, for some $v \in L_\infty(T)$ with $\|v\|_\infty < \delta/2$ and $v(t) = 0$ almost everywhere on T_1^c.

Finally, set $y = x_m - v$. Then $y \in L_\infty(T)$, and almost everywhere on T_1^c, $y(t) = x_m(t) \geqq 1/m$. On T_1 we know $x(t) \geqq \delta > 1/m$ almost everywhere, so $x_m(t) \geqq \min\{x(t), m\} \geqq \delta$, almost everywhere. Since $v(t) \leqq \delta/2$ almost everywhere it follows that $y(t) \geqq \delta/2$ almost everywhere on T_1. The result now follows. □

For a given set of constraint functions a_i, $i = 1, \cdots, n$, the question of for what b's the problem (EP_p) is consistent is known as the *extendibility problem.* Simple conditions are known, for example, when $T = [0, 1]$ with Lebesgue measure and $a_i(t) = t^{i-1}$ (one of the "classical" moment problems), and for the case when $T = [-\pi, \pi]$ with Lebesgue measure and $\{a_1(\theta), \cdots, a_{2k+1}(\theta)\} = \{1, \cos\theta, \sin\theta, \cdots, \cos k\theta, \sin k\theta\}$ (the trigonometric moment problem). See, for example, [Karlin and Studden, 1966] and [Ben-Tal, Borwein, and Teboulle, 1988b] for more information.

PROPOSITION 2.11. *If ϕ_+ is strictly convex on* dom ϕ_+, *then any optimal solution to* (EP$_p$) *is unique.*

Proof. If ϕ_+ is strictly convex on dom ϕ_+ then so is I_{ϕ_+} on dom I_{ϕ_+}. To see this, suppose $x_1, x_2 \in \operatorname{dom} I_{\phi_+}$, with $x_1 \neq x_2$, and $0<\gamma<1$. By strict convexity of ϕ_+, $\gamma\phi_+(x_1(t))+(1-\gamma)\phi_+(x_2(t)) \leqq \phi_+(\gamma x_1(t)+(1-\gamma)x_2(t))$, with strict inequality on a nonnull set. It follows that $\gamma I_{\phi_+}(x_1)+(1-\gamma)I_{\phi_+}(x_2)<I_{\phi_+}(\gamma x_1+(1-\gamma)x_2)$. The result now follows. □

3. Examples. In the course of the derivation of a primal solution described in § 1, the crucial step was differentiating through the integral in the dual objective function to obtain (1.1). The following example shows clearly the sort of difficulties that can occur.

For $1 \leqq p<2$, consider the convex integral functional $I: L_p[0,1] \to (-\infty, \infty]$ defined by

$$I(x)=\begin{cases} -\int_0^1 \log x(t)\, dt, & x(t)>0 \text{ a.e.}, \\ +\infty, & \text{otherwise}. \end{cases}$$

By [Rockafellar, Thm. 22, 1974], $y \in \partial I(x)$ if and only if $y(t) \in \partial(-\log)(x(t))=\{-1/x(t)\}$, almost everywhere, for $x(t)>0$ almost everywhere. With a slight abuse of notation, let us denote the function that is identically equal to one by 1. Then $\partial I(1)=\{-1\}$, so I has a unique subgradient at 1.

Now define $h \in L_p[0,1]$ by $h(t)=1/\sqrt{t}$. For any $\varepsilon>0$, $1-\varepsilon/\sqrt{t} \leqq 0$ on a set of nonzero measure, so $I(1-(\varepsilon/\sqrt{t}))=+\infty$. Thus, the directional derivative $I'(1;-h)=+\infty$.

On the other hand, we have

$$I'(1;h)=\lim_{\varepsilon \downarrow 0} \frac{1}{\varepsilon}\left\{-\int_0^1 \log\left(1+\frac{\varepsilon}{\sqrt{t}}\right) dt\right\}.$$

It is straightforward to check that as $\varepsilon \downarrow 0$, $(1/\varepsilon)\log(1+(\varepsilon/\sqrt{t})) \uparrow (1/\sqrt{t})$, so by the monotone convergence theorem, $I'(1;h)=-\int_0^1 (1/\sqrt{t})\, dt=-2$.

Thus, I has no Gateaux derivative at 1 although it has a unique subgradient there. The choice of space is clearly important here. In the case $p=\infty$, we have $1 \in \|\cdot\|_\infty-\operatorname{int}(\operatorname{dom} I)$, and it is easy to see that I is $\|\cdot\|_\infty$-continuous at 1. Thus, in this case the Gateaux derivative does exist: $\nabla I(1)=-1$ (see, for example, [Holmes, 1975]).

Simple examples show that even when the constraint qualification is satisfied we cannot necessarily expect primal attainment in (EP$_p$).

Example 3.1. Consider the following semi-infinite linear program:

$$\begin{array}{ll} \inf & \int_0^1 x(t)\, dt \\ \text{subject to} & \int_0^1 t x(t)\, dt=1, \qquad x \geqq 0, \quad x \in L_p[0,1]. \end{array}$$

In this problem $\phi(x)=x$, so $(\phi_+)^*(y)=\sup_{x \geqq 0}\{xy-x\}=\delta(y \mid (-\infty,1])$. Thus, the dual problem is

$$\begin{array}{ll} \max & \lambda-\int_0^1 \delta(\lambda t \mid (-\infty,1])\, dt \\ \text{subject to} & \lambda \in \mathbb{R}. \end{array}$$

The optimal dual solution is clearly $\bar{\lambda}=1$, giving value 1. The constraint qualification is clearly satisfied by $\hat{x}(t)=2$, so the primal value is also 1, by Corollary 2.6. However, since $x(t) > tx(t)$ almost everywhere, $\int_0^1 x(t)\,dt > \int_0^1 tx(t)\,dt = 1$ for any feasible x, so the optimal value is not attained (for any $1 \leqq p \leqq \infty$).

The above behavior can be observed even when the function ϕ is essentially strictly convex.

Example 3.2. Consider the following problem:

$$\begin{aligned} &\inf && \int_0^{2\pi} \left(\frac{1}{x(t)}\right) dt \\ &\text{subject to} && \int_0^{2\pi} x(t)\sin t\,dt = 1, \qquad x \geqq 0, \quad x \in L_p[0, 2\pi]. \end{aligned}$$

In this case $\phi(x) = 1/x$, $x > 0$, and ∞ otherwise. Thus, $(\phi_+)^*(y) = \sup_{x>0}\{xy - (1/x)\}$, so by differentiation,

$$(\phi_+)^*(y) = \begin{cases} -2\sqrt{-y}, & y \leqq 0, \\ \infty, & \text{otherwise.} \end{cases}$$

The dual problem therefore becomes

$$\begin{aligned} &\max && \lambda - \int_0^{2\pi} (-2\sqrt{-\lambda \sin t})\,dt \\ &\text{subject to} && -\lambda \sin t \geqq 0 \text{ a.e. on } [0, 2\pi], \qquad \lambda \in \mathbb{R}. \end{aligned}$$

The only dual feasible solution is $\bar{\lambda} = 0$, so the dual value is 0. The constraint qualification is satisfied by

$$\hat{x}(t) = \begin{cases} 1, & 0 \leqq t < \pi \\ \frac{1}{2}, & \pi \leqq t \leqq 2\pi, \end{cases}$$

since

$$\int_0^{2\pi} \hat{x}(t)\sin t\,dt = [-\cos\theta]_0^{\pi} + \frac{1}{2}[-\cos\theta]_{\pi}^{2\pi} = 1.$$

Thus, the primal value is also 0, by Corollary 2.6, or direct computation. However, this value clearly is not attained.

In both of the above examples the function $(\phi_+)^*$ is differentiable on the interior of its domain, but the manipulation described in the introduction is manifestly invalid since there exists no primal optimal solution. Our last example shows how attainment in the primal problem can depend on the choice of space for the primal variables.

Example 3.3. Consider the following problem:

$$\begin{aligned} &\inf && \int_0^1 \left(\frac{1}{x(t)}\right) dt \\ &\text{subject to} && \int_0^1 tx(t)\,dt = 1, \qquad x \geqq 0, \quad x \in L_p[0, 1]. \end{aligned}$$

In a similar fashion to Example 3.2, the dual problem is

$$\begin{aligned} &\max && \lambda - \int_0^1 (-2\sqrt{-\lambda t})\,dt \\ &\text{subject to} && -\lambda t \geqq 0, \text{ a.e. on } [0, 1], \qquad \lambda \in \mathbb{R}. \end{aligned}$$

This is equivalent to $\max\{\lambda+\frac{4}{3}\sqrt{-\lambda}\,|\,\lambda\leqq 0\}$. By differentiation, the optimum occurs at $\bar{\lambda}=-\frac{4}{9}$, giving a dual value of $\frac{4}{9}$.

The constraint qualification is again satisfied by, for instance, $\hat{x}(t)=2$, almost everywhere on $[0,1]$. The primal value is therefore also $\frac{4}{9}$. The manipulation described in the introduction (differentiating the dual objective function) may now be performed, at least formally, to obtain $\bar{x}(t)=3/2\sqrt{t}$. It may be checked that $\bar{x}$ satisfies the primal constraint and has value $\frac{4}{9}$. Thus, it is optimal, at least for $1\leqq p<2$, and by Proposition 2.11 it is the unique primal optimal solution. However, $\bar{x}\notin L_p[0,1]$ for $2\leqq p\leqq\infty$, so for these values of p there exists no primal optimal solution (otherwise we would obtain a contradiction to Proposition 2.11). Note finally that the question of primal attainment for this problem is not affected by the addition of the extra constraint $\int_0^1 x(t)\,dt=3$, since $\bar{x}$ remains feasible. This provides a counterexample whose constraints are in the form of a standard moment problem.

4. Primal solutions. We shall now give a rigorous treatment of conditions ensuring the existence of a primal optimal solution. We begin with some simple results aimed at identifying the domain of the functional $I_{\phi_+^*}$.

PROPOSITION 4.1. *The function* $(\phi_+)^*:\mathbb{R}\to(-\infty,\infty]$ *is monotone increasing.*

Proof. The result follows immediately from the definition:

$$(\phi_+)^*(y)=\sup_{x\geqq 0}\{yx-\phi(x)\}. \qquad \square$$

LEMMA 4.2. *Suppose* $\phi:\mathbb{R}\to(-\infty,\infty]$ *with* ϕ_+ *closed, convex, and proper. Define* $d:=\lim_{x\to\infty}(\phi(x)/x)$ (*this limit exists*). *Then* $\operatorname{cl}(\operatorname{dom}(\phi_+)^*)=(-\infty,d]$.

Proof. Define $\psi(x)=\phi_+(x)-y(x)$. Then by [Rockafellar, Cor. 13.3.4, 1970], $y\in\operatorname{cl}(\operatorname{dom}(\phi_+)^*)$ if and only if $(\psi 0^+)(z)\geqq 0$, for all z. By [Rockafellar, Thm. 8.5, 1970],

$$\begin{aligned}(\psi 0^+)(z)&=\lim_{\lambda\to\infty}\frac{\psi(v+\lambda z)-\psi(v)}{\lambda}\\&=\lim_{\lambda\to\infty}\frac{\phi_+(v+\lambda z)-\phi_+(v)-y\lambda z}{\lambda}\\&=\lim_{\lambda\to\infty}\left(\frac{\phi_+(v+\lambda z)}{v+\lambda z}\,\frac{v+\lambda z}{\lambda}\right)-yz\\&=(d-y)z,\end{aligned}$$

for any $v\in\operatorname{dom}\phi_+$ and for $z>0$. For $z<0$, $(\psi 0^+)(z)=+\infty$, and $(\psi 0^+)(0)=0$. Thus, $(\psi 0^+)(z)\geqq 0$, for all z, if and only if $y\leqq d$. The result now follows by Proposition 4.1. $\square$

LEMMA 4.3. *Suppose* ϕ *is as in Lemma* 4.2. *The functional* $I_{(\phi_+)^*}:L_\infty(T)\to(-\infty,\infty]$ *is* $\|\cdot\|_\infty$*-continuous at any* $y\in L_\infty(T)$ *for which* $\operatorname{ess\,sup} y<d$.

Proof. There exist $M>0$ and $\varepsilon>0$ such that $-M\leqq y(t)\leqq d-\varepsilon$, almost everywhere on T. By Lemma 4.2, $\operatorname{int}(\operatorname{dom}(\phi_+)^*)=(-\infty,d)$, so $(\phi_+)^*$ is continuous on this set, by [Rockafellar, Thm. 10.1, 1970]. It follows that $(\phi_+)^*$ is uniformly continuous on $[-2M,d-(\varepsilon/2)]$. Thus, if $y_n\to y$ uniformly on T, $(\phi_+)^*(y_n(t))\to(\phi_+)^*y(t)$ uniformly, and so $I_{\phi_+^*}(y_n)\to I_{\phi_+^*}(y)$, as required. Alternatively, it is sufficient to observe that $I_{\phi_+^*}$ is bounded above on some neighborhood of y [Holmes, 1975]. $\square$

We will use the following ideas from [Rockafellar, 1970].

DEFINITION 4.4. A proper convex function $f:\mathbb{R}^n\to(-\infty,\infty]$ is *essentially strictly convex* if f is strictly convex on every convex subset of $\{x\,|\,\partial f(x)\neq\phi\}$.

Note that if $n=1$, f is essentially strictly convex if and only if f is strictly convex on dom f. This follows from the fact that if $f:\mathbb{R}^n\to(-\infty,\infty]$ is convex, the only way in which it can fail to be strictly convex on dom f is if it is actually affine on some line segment. To see this, suppose f is not strictly convex on dom f. Without loss of generality suppose $f(0)=f(x_0)=0$ for some $x_0\neq 0$, and that for some $0<\lambda_0<1$, $f(\lambda_0x_0)=0$. Convexity implies $f(\lambda x_0)\leqq 0$ for all $0\leqq\lambda\leqq 1$, so suppose we have strict inequality for some λ, and without loss of generality suppose $0<\lambda<\lambda_0$. We then have

$$\begin{aligned}0&=f(\lambda_0x_0)\\&=f\left(\left(\frac{1-\lambda_0}{1-\lambda}\right)(\lambda x_0)+\left(\frac{\lambda_0-\lambda}{1-\lambda}\right)(x_0)\right)\\&\leqq\left(\frac{1-\lambda_0}{1-\lambda}\right)f(\lambda x_0)+\left(\frac{\lambda_0-\lambda}{1-\lambda}\right)f(x_0)\\&<0,\end{aligned}$$

which is a contradiction. Thus, $f(x)=0$ on $[0,x_0]$.

However, if f is essentially strictly convex then it is certainly strictly convex on ri (dom f) $\subset\{x\,|\,\partial f(x)\neq\phi\}$, and the above argument then shows it must be strictly convex on dom f (provided $n=1$).

DEFINITION 4.5. A proper convex function $f:\mathbb{R}^n\to(-\infty,\infty]$ is *essentially smooth* if f is differentiable on int (dom f) $\neq\phi$, and $\|\nabla f(x_i)\|\to\infty$ whenever $(x_i)\subset$ int (dom f) with $x_i\to$ some x in the boundary of dom f.

THEOREM 4.6. *If $f:\mathbb{R}^n\to(-\infty,\infty]$ is closed, proper, and convex, then f is essentially strictly convex if and only if f^* is essentially smooth.*

Proof. For the proof see [Rockafellar, Thm. 26.3, 1970]. □

The following result gives conditions for dual solutions to be unique.

THEOREM 4.7. *Consider the problem* (DEP$_p$) *of Corollary* 2.6. *Suppose* $\{a_1,\cdots,a_n\}$ *is linearly independent and ϕ_+ is essentially smooth. Then any optimal solution is unique.*

Proof. Suppose $\lambda^1\neq\lambda^2$ are both optimal solutions of (DEP$_p$). The two functions $\sum_{i=1}^n\lambda_i^ja_i(t)\in$ dom $(\phi_+)^*$ almost everywhere, $j=1,2$. Since the a_i's are linearly independent, they differ on a nonnull subset of T, T_1 say. By Theorem 4.5, $(\phi_+)^*$ is strictly convex on dom $(\phi_+)^*$, so if we set $\lambda^3=\frac{1}{2}\lambda^1+\frac{1}{2}\lambda^2$, then

$$(\phi_+)^*\left(\sum_{i=1}^n\lambda_i^3a_i(t)\right)\leqq\frac{1}{2}(\phi_+)^*\left(\sum_{i=1}^n\lambda_i^1a_i(t)\right)+\frac{1}{2}(\phi_+)^*\left(\sum_{i=1}^n\lambda_i^2a_i(t)\right),$$

almost everywhere, with strict inequality on T_1. It follows that λ^3 is an improvement on λ^1 and λ^2, which is a contradiction. □

We are now ready to prove our main result. We shall be concerned specifically with the dual pair (EP$_1$) and (DEP$_1$), so the primal variable will lie in $L_1(T)$. In this result we shall give conditions allowing us to obtain the primal optimal solution by differentiating the dual objective function at the optimum, as described in the introduction. This function involves the convex functional $I_{\phi_+^*}:L_\infty(T)\to(-\infty,\infty]$, whose subgradients lie in $L_\infty^*(T)$. We shall use the results in [Rockafellar, 1971] to decompose such subgradients into their singular and continuous (lying in $L_1(T)$) parts. Finally, we shall find a condition ensuring the singular part vanishes, which will lead to the desired conclusion. The result could alternatively be proved by a direct differentiation argument (see [Ben-Tal, Borwein, and Teboulle, 1988b]), but we shall use the ideas from our proof again in the last section.

THEOREM 4.8. *Consider the dual pair of problems* (EP_1) *and* (DEP_1). *Suppose that all the assumptions of Corollary 2.6 are met, and that* ϕ_+ *is strictly convex on* dom ϕ_+. *Let* $\bar{\lambda}$ *be dual optimal. Suppose finally that the following assumption is satisfied*:

$$d := \lim_{x\to\infty} \frac{\phi(x)}{x} > \operatorname{ess\,sup} \sum_{i=1}^{n} \bar{\lambda}_i a_i. \tag{4.9}$$

Then the unique primal optimal solution is given by

$$\bar{x}(t) = ((\phi_+)^*)'\left(\sum_{i=1}^{n} \bar{\lambda}_i a_i(t)\right).$$

Proof. By Corollary 2.6 we know that the primal and dual values are equal, with dual attainment. Let us denote the constraint map by $A: L_1(T) \to \mathbb{R}^n$, so $(Ax)_i = \langle x, a_i\rangle$, each i. The adjoint map $A^T: \mathbb{R}^n \to L_\infty(T)$ is therefore defined by $A^T\lambda = \sum_{i=1}^n \lambda_i a_i$. It is thus $\mathbb{R}^n - \|\cdot\|_\infty$ continuous, so we can define $A^{TT}: L_\infty^*(T) \to \mathbb{R}^n$, and we will have $A^{TT}|_{L_1(T)} = A$.

The dual objective function is $-g(\lambda)$, where $g: \mathbb{R}^n \to (-\infty, \infty]$ is defined by $g(\lambda) = -b^T\lambda + I_{\phi_+^*}(A^T\lambda)$, and since $\bar{\lambda}$ is dual optimal, $0 \in \partial g(\bar{\lambda})$. However, $\sum_{i=1}^n \bar{\lambda}_i a_i \in$ range (A^T) and $I_{\phi_+^*}$ is continuous at this point, by (4.9) and Lemma 4.3. It follows by [Borwein, Thm. 4.1, 1981] or [Rockafellar, Thm. 19, 1974] that $\partial g(\bar{\lambda}) = -b + A^{TT}\,\partial I_{\phi_+^*}(A^T\bar{\lambda})$, and so there exists $\bar{\mu} \in \partial I_{\phi_+^*}(A^T\bar{\lambda}) \subset L_\infty^*(T)$ with $A^{TT}\bar{\mu} = b$.

By definition of the subgradient, for all $y \in L_\infty(T)$, $\bar{\mu}(y - A^T\bar{\lambda}) \leqq I_{\phi_+^*}(y) - I_{\phi_+^*}(A^T\bar{\lambda})$. Thus, $\bar{\mu} \geqq 0$, since if not, we could find a $y \leqq A^T\bar{\lambda}$ with $\bar{\mu}(y - A^T\bar{\lambda}) > 0$, implying $I_{\phi_+^*}(A^T\bar{\lambda}) < I_{\phi_+^*}(y)$, which contradicts the fact that $(\phi_+)^*$ is monotone increasing by Proposition 4.1.

We can now apply [Rockafellar, Cor. 1B, 1971] to deduce the existence of a $z \in L_1(T)$ and a singular component $\nu \in L_\infty^*(T)$ such that $\bar{\mu} = z + \nu$, where $z(t) \in \partial(\phi_+)^*((A^T\bar{\lambda})(t))$, almost everywhere and ν attains its maximum over dom $I_{\phi_+^*}$ at $A^T\bar{\lambda}$. However, we know $A^T\bar{\lambda} \in$ int (dom $I_{\phi_+^*}$), from which it follows that $\nu = 0$ (alternatively, see [Rockafellar, Cor. 2C, 1971]). Furthermore, we know $(A^T\bar{\lambda})(t) \in$ int (dom $(\phi_+)^*$), almost everywhere, and from Theorem 4.6 we know $(\phi_+)^*$ is essentially smooth. Thus, $z \in L_1(T)$, $z \geqq 0$, $Az = b$, and $z(t) = ((\phi_+)^*)'((A^T\bar{\lambda})(t))$, almost everywhere. In particular, z is primal feasible.

Finally, since ϕ_+ is closed,

$$\phi_+(z(t)) + (\phi_+)^*((A^T\bar{\lambda})(t)) = z(t)(A^T\bar{\lambda})(t), \quad \text{a.e.,}$$

by [Rockafellar, Thm. 23.5, 1970]. Integrating over T gives

$$\begin{aligned} I_{\phi_+}(z) + I_{\phi_+^*}(A^T\bar{\lambda}) &= \langle z, A^T\bar{\lambda}\rangle \\ &= (Az)^T\bar{\lambda} \\ &= b^T\bar{\lambda}, \end{aligned}$$

so z has the same objective value as the dual value, so it is optimal by weak duality. Uniqueness follows by Proposition 2.11. □

5. Special cases. The last result of the previous section (Theorem 4.8) showed that under suitable conditions, if the dual optimal solution $\bar{\lambda}$ satisfied the condition that ess sup $\sum_{i=1}^n \bar{\lambda}_i a_i < d$, where d was defined to be $\lim_{x\to\infty}(\phi(x)/x)$, we could obtain the unique primal solution by differentiating the dual objective function. If $d = \infty$ then this condition is clearly no restriction. However, if $d < \infty$ then condition (4.9) may fail; this is the case, for instance, in Example 3.2. In this final section we shall consider conditions on ϕ that ensure a priori that condition (4.9) will hold.

We begin by defining another constant associated with ϕ. Assuming ϕ_+ is essentially strictly convex, we know by Lemma 4.2 that $\mathrm{cl}\,(\mathrm{dom}\,(\phi_+)^*)=(-\infty, d]$, and that $(\phi_+)^*$ is essentially smooth by Theorem 4.6. Let us define (assuming $d<\infty$)

$$c:=\underline{\lim}_{y\uparrow d}\,(d-y)((\phi_+)^*)'(y). \tag{5.1}$$

Lemma 5.2. *Suppose ϕ_+ is essentially strictly convex and essentially smooth. Then*

$$c=\underline{\lim}_{x\to\infty}\,(d-\phi'_+(x))x. \tag{5.3}$$

Proof. By [Rockafellar, Thm. 26.5, 1970], ϕ'_+ is one-to-one from $\mathrm{int}\,(\mathrm{dom}\,\phi_+)$ to $\mathrm{int}\,(\mathrm{dom}\,(\phi_+)^*)=(-\infty, d)$, continuous in both directions, and $((\phi_+)^*)'=(\phi'_+)^{-1}$. Note that since $d<\infty$, $(k,\infty)\subset \mathrm{dom}\,\phi_+$, for k sufficiently large, so expression (5.3) is well defined. Furthermore, $\phi'_+(x)$ increases to d as x increases to ∞, continuously, using l'Hôpital's rule and the convexity of ϕ_+. Thus, we have

$$\begin{aligned} c&=\underline{\lim}_{y\uparrow d}\,(d-y)((\phi_+)^*)'(y)\\ &=\underline{\lim}_{x\uparrow\infty}\,(d-\phi'_+(x))((\phi_+)^*)'(\phi'_+(x))\\ &=\underline{\lim}_{x\to\infty}\,(d-\phi'_+(x))x, \end{aligned}$$

as required. □

We will now restrict our attention to the case where the underlying measure space T is a compact real interval with Lebesgue measure. We will need the following lemma.

Lemma 5.4. *Suppose $h\in C[\alpha,\beta]$, $h(t_0)=0$ for some $\alpha\leqq t_0\leqq\beta$, and h is Lipschitz (or in particular, continuously differentiable) at t_0. Then $1/h\notin L_1[\alpha,\beta]$.*

Proof. Suppose first that h is continuously differentiable at t_0. Define a function $g:[\alpha,\beta]\to\mathbb{R}$ by

$$g(t)=\begin{cases} h(t)/(t-t_0), & t\neq t_0,\\ h'(t_0), & t=t_0.\end{cases}$$

By L'Hôpital's rule, g is continuous, so by compactness, $|g(t)|\leqq M$ for all $t\in[\alpha,\beta]$, for some M. Since $h(t)=(t-t_0)g(t)$ for all $t\in[\alpha,\beta]$, $|h(t)|\leqq M|t-t_0|$, so h is Lipschitz at t_0.

Now assume h is Lipschitz at t_0, so $|h(t)|\leqq M|t-t_0|$, for all $t\in[\alpha,\beta]$. Thus,

$$\frac{1}{|h(t)|}\geqq\frac{1}{M|t-t_0|}\quad \forall t\in[\alpha,\beta],\quad t\neq t_0.$$

Since clearly $1/|t-t_0|\notin L_1[\alpha,\beta]$, the result follows. □

For the last step in the above argument, it is critical that the underlying measure is Lebesgue measure (or at least is greater than some positive multiple of Lebesgue measure). The underlying space is important, too. Consider, for example, $T=[-1,1]\times[-1,1]\subset\mathbb{R}^2$, with Lebesgue measure. Then $1/\|t\|\in L_1(T)$.

Theorem 5.5. *Let $T=[\alpha,\beta]$, with Lebesgue measure, and each a_i be locally Lipschitz (or in particular, continuously differentiable), $i=1,\cdots,n$. Consider the dual pair of problems* (EP$_1$) *and* (DEP$_1$). *Suppose that all of the assumptions of Corollary* 2.6 *are met, and that ϕ_+ is strictly convex on* dom ϕ_+. *Suppose that*

$$\textit{there exists } \mu\in\mathbb{R}^n \textit{ with } \sum_{i=1}^{n}\mu_i a_i(s)<d\quad \forall s\in[\alpha,\beta], \tag{5.6}$$

where $d := \lim_{x\to\infty} (\phi(x)/x)$, and if $d<\infty$ define $c := \lim_{y\uparrow d} (d-y)((\phi_+)^*)'(y)$. Suppose that either $d=\infty$ or $d<\infty$ and $c>0$. Then the unique primal optimal solution is given by $\bar{x}(t) = ((\phi_+)^*)'(\sum_{i=1}^n \bar{\lambda}_i a_i(t))$, where $\bar{\lambda}$ is a dual optimal solution.

Proof. Just as in the proof of Theorem 4.8, we know that $0 \in \partial g(\bar{\lambda})$, where $g(\lambda) = -b^T\lambda + I_{\phi_+^*}(A^T\lambda)$. We can write (5.6) as

$$\text{range}\,(A^T) \cap \text{cont}\,(I_{\phi_+^*}) \neq \phi,$$

by Lemma 4.3, so it follows by [Rockafellar, Thm. 19, 1974] or [Borwein, Thm. 4.1, 1981] that $\partial g(\lambda) = -b + A^{TT}\,\partial I_{\phi_+^*}(A^T\lambda)$. Thus, there exists $\bar{\mu} \in \partial I_{\phi_+^*}(A^T\bar{\lambda})$. Applying [Rockafellar, Cor. 1B, 1971] just as in the proof of Theorem 4.8, it follows that there exists $z \in L_1(T)$ with $z(t) \in \partial(\phi_+)^*((A^T\bar{\lambda})(t))$, almost everywhere.

The next step is to observe that since $(\phi_+)^*$ is essentially smooth (by Theorem 4.6), $\partial(\phi_+)^*(d) = \phi$. To see this, recall that as $y \uparrow d$, $((\phi_+)^*)'(y) \uparrow \infty$, by the definition of essentially smooth, and the fact that, by Lemma 4.2, $\text{cl}\,(\text{dom}\,(\phi_+)^*) = (-\infty, d]$. If $u \in \partial(\phi_+)^*(d)$ then $u(y-d) \leqq (\phi_+)^*(y) - (\phi_+)^*(d)$, for all y. By the mean value theorem, given any $M>0$ there exists $y<d$ with $(\phi_+)^*(d) - (\phi_+)^*(y) = ((\phi_+)^*)'(z)(d-y) \geqq M(d-y)$, for some $y<z<d$. It follows that $u \geqq M$, which is a contradiction as M was arbitrary.

Since $z(t) \in \partial(\phi_+)^*((A^T\bar{\lambda})(t))$ almost everywhere, it follows that $(A^T\bar{\lambda})(t) < d$, almost everywhere, and so $z(t) = ((\phi_+)^*)'((A^T\bar{\lambda})(t))$, almost everywhere. We know from Theorem 4.8 that if $\text{ess sup}\, A^T\bar{\lambda} < d$ then $\bar{x} = z$ is the unique primal optimal solution, as required. Since the a_i's are continuous on $[\alpha, \beta]$, the only alternative is $(A^T\bar{\lambda})(t_0) = d$, for some $\alpha \leqq t_0 \leqq \beta$. We will show that, in this case, z cannot possibly lie in $L_1[\alpha, \beta]$, giving a contradiction. This will complete the proof.

Assume therefore that for some $\alpha \leqq t_0 \leqq \beta$, $\sum_{i=1}^n \bar{\lambda}_i a_i(t_0) = d < \infty$. By the definition of c, there exists $\varepsilon > 0$ such that $(d-y)((\phi_+)^*)'(y) \geqq c/2 > 0$, for all $d > y \geqq d - \varepsilon$. By the continuity of the a_i's, there exists $\delta > 0$ such that $\sum_{i=1}^n \bar{\lambda}_i a_i(t) \geqq d - \varepsilon$, for all $|t - t_0| \leqq \delta$. Thus,

$$\begin{aligned} z(t) &= ((\phi_+)^*)'\left(\sum_{i=1}^n \bar{\lambda}_i a_i(t)\right) \\ &\geqq \frac{c}{2(d - \sum_{i=1}^n \bar{\lambda}_i a_i(t))}, \end{aligned}$$

for all $|t - t_0| \leqq \delta$. But by Lemma 5.4, the right-hand side is not integrable (on either $[t_0 - \delta, t_0]$ or $[t_0, t_0 + \delta]$), so $z \notin L_1[\alpha, \beta]$. This completes the proof. □

Note in particular that condition (5.6) will always hold if one of the a_i's is a nonzero constant function.

We will conclude with a number of examples of typical objective functions ϕ, taken from [Ben-Tal, Borwein, and Teboulle, 1988b]. Some of these objectives are taken from the literature, others are new. For each function we give the numbers d and c of Theorem 5.5. It is easy to check that ϕ_+ is closed and essentially strictly convex in each case.

Examples 5.6.

(i) $$\phi_+(x) = \begin{cases} -\log x, & x>0, \\ \infty, & x \leqq 0, \end{cases}$$

$$(\phi_+)^*(y) = \begin{cases} -1-\log(-y), & y<0, \\ \infty, & y \geqq 0, \end{cases} \qquad d=0, \quad c=1.$$

(ii) $$\phi_+(x)=\begin{cases} x\log x - x, & x>0,\\ 0, & x=0,\\ \infty, & x<0,\end{cases}$$

$$(\phi_+)^*(y)=e^y, \qquad d=\infty.$$

(iii) $$\phi_+ + (x)=\begin{cases} \dfrac{1}{p}x^p, & x\geqq 0, \quad (\text{with } 1<p<\infty),\\ \infty, & x<0,\end{cases}$$

$$(\phi_+)^*(y)=\frac{1}{q}(y^+)^q, \qquad d=\infty.$$

(iv) $$\phi_+(x)=\begin{cases} e^x-1, & x\geqq 0,\\ \infty, & x<0,\end{cases}$$

$$(\phi_+)^*(y)=\begin{cases} y\log y - y+1, & y\geqq 1,\\ 0, & y<1,\end{cases} \qquad d=\infty.$$

(v) $$\phi_+(x)=\begin{cases} \dfrac{\gamma x - x^\gamma}{1-\gamma}, & x\geqq 0,\\ \infty, & x<0,\end{cases} \quad \left(\text{with } 0<\gamma<1,\ \frac{1}{\gamma}-\frac{1}{\psi}=1\right),$$

$$(\phi_+)^*(y)=\begin{cases} \left(1-\dfrac{1}{\psi}y\right)^{-\psi}, & y\leq \psi,\\ \infty, & \text{otherwise},\end{cases} \qquad d=\psi, \quad c=\infty.$$

(vi) $$\phi_+(x)=\begin{cases} k-\sqrt{k^2-x^2}, & 0\leqq x\leqq k,\\ \infty, & \text{otherwise},\end{cases}$$

$$(\phi_+)^*(y)=\begin{cases} k(\sqrt{1+y^2}-1), & y\geqq 0,\\ 0, & y<0,\end{cases} \qquad d=\infty.$$

(vii) $$\phi_+(x)=\begin{cases} x\log x-(1+x)\log(1+x), & x>0,\\ 0, & x=0,\\ \infty, & x<0,\end{cases}$$

$$(\phi_+)^*(y)=\begin{cases} -\log(1-e^y), & y<0,\\ \infty, & y\geqq 0,\end{cases} \qquad d=0, \quad c=\infty.$$

(viii) $$\phi_+(x)=\begin{cases} \dfrac{1}{x}, & x>0,\\ \infty, & x\leqq 0,\end{cases}$$

$$(\phi_+)^*(y)=\begin{cases} -2\sqrt{-y}, & y\leqq 0,\\ \infty, & y>0,\end{cases} \qquad d=0, \quad c=0.$$

In all of the above examples, ϕ_+ is strictly convex on its domain, and except in cases (iii), (iv), and (vi) it is also essentially smooth. With the exception of case (vi), $(0,\infty)\subset \operatorname{dom}\phi_+$, so the constraint qualification will simply require the existence of a feasible $\hat{x}\in\operatorname{dom} I_{\phi_+}$ with $\hat{x}(t)>0$, almost everywhere. In the case of (vi) it can be checked using the results of [Borwein and Lewis, 1988] that the constraint qualification requires the existence of a feasible $\hat{x}$ with $0<\hat{x}(t)<k$, almost everywhere. Example (viii) shows, as we would expect, that the assumptions of Theorem 5.5 fail for Example 3.2.

Suppose finally that $(\phi_+)^*$ is actually continuously differentiable on (∞, d), as is the case in all of the above examples. With the assumptions of Theorem 5.5 it then follows that the unique optimal solution of (EP_1), $\bar{x}$, is actually continuous on T, so it certainly lies in L_p for any $1 \leqq p \leqq \infty$. Thus, $\bar{x}$ will in fact be the unique optimal solution for (EP_p), for any $1 \leqq p \leqq \infty$, and also for the analogous problem posed in $C[\alpha, \beta]$.

REFERENCES

A. Ben-Tal, J. M. Borwein, and M. Teboulle (1988a), *A dual approach to multidimensional L_p spectral estimation problems*, SIAM J. Control Optim., 26, pp. 985-996.

——— (1988b), *Spectral estimation via convex programming*, to appear.

S. Bochner and W. T. Martin (1964), *Several Complex Variables*, Princeton University Press, Princeton, NJ.

J. M. Borwein (1981), *A Lagrange multiplier theorem and a sandwich theorem for convex relations*, Math. Scand., 48, pp. 188-204.

J. M. Borwein and A. S. Lewis (1988), *Partially finite convex programming, Parts* I *and* II, Math. Programming, to appear.

J. P. Burg (1975), *Maximum entropy spectral analysis*, Ph.D. dissertation, Stanford University, Stanford, CA.

B. K. Goodrich and A. Steinhardt (1986), *L_2 spectral estimation*, SIAM J. Appl. Math., 46, pp. 417-428.

B. B. Holmes (1975), *Geometric Functional Analysis and its Applications*, Springer-Verlag, New York.

A. D. Ioffe and V. M. Tihomorov (1968), *Extension of variational problems*, Trans. Moscow Math. Soc., 18, pp. 207-273.

L. D. Irvine, S. P. Marin, and P. W. Smith (1986), *Constrained interpolation and smoothing*, Constructive Approximation, 2, pp. 129-151.

R. W. Johnson and J. E. Shore (1984), *Which is the better entropy expression for speech processing: slogs or logs?*, IEEE Trans. on Acoustics, Speech and Signal Processing, ASSP 32, pp. 129-136.

S. Karlin and W. Studden (1966), *Tchebycheff systems with applications in analysis and statistics*, Interscience, New York.

C. A. Micchelli, P. W. Smith, J. Swetits, and J. D. Ward (1985), *Constrained L_p approximation*, Constructive Approximation, 1, pp. 93-102.

R. T. Rockafellar (1968), *Integrals which are convex functionals*, Pacific J. Math., 24, pp. 525-539.

——— (1970), *Convex Analysis*, Princeton University Press, Princeton, NJ.

——— (1971), *Integrals which are convex functionals* II, Pacific J. Math., 39, pp. 439-469.

——— (1974), *Conjugate Duality and Optimization*, CBMS-NSF Regional Conference Series in Applied Mathematics 16, Society for Industrial and Applied Mathematics, Philadelphia, PA.

——— (1976), *Integral functions, normal integrands and measurable selections*, in Nonlinear Operators and the Calculus of Variations, L. Waelbroeck, ed., Lecture Notes in Math. 543, Springer-Verlag, New York, Berlin, pp. 157-207.

W. Rudin (1966), *Real and Complex Analysis*, McGraw-Hill, New York.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 339-346, March 1991

005

1-DETERMINACY OF FEASIBLE SETS*

TOSHIHIRO MATSUMOTO†, SUSUMU SHINDOH‡, AND RYUICHI HIRABAYASHI§

Abstract. Germs of feasible sets, 1-determinacy of feasible sets, (strongly) restricted A-equivalence of map germs, and (strongly) restricted 1-determinacy of map germs are defined. Then each concept of 1-determinacy is characterized by a necessary and sufficient condition. Finally, it is demonstrated that all of these concepts are equivalent in some sense.

Key words. feasible sets, 1-determinacy, (strongly) restricted A-equivalence, singularity

1. Introduction.

There are many articles on the study of nonlinear programs from the topological point of view (see, for example, Kojima [9], Jongen, Jonker, and Twilt [7], [8], Guddat and Jongen [4], Fujiwara [1], etc.). In 1986, Guddat, Jongen, and Rueckmann [5] proved that a feasible set of a nonlinear program becomes a topological manifold under the assumption that Mangasarian–Fromovitz constraint qualification (MF-condition) holds at every point of the feasible set. Guddat, Jongen, and Rueckmann [5], Jongen, Jonker, and Twilt [8], and Guddat and Jongen [4] developed a structural stability theory of nonlinear programs. Especially, in [5], Guddat, Jongen, and Rueckmann proved that it is necessary and sufficient to assume the condition mentioned above for a feasible set to be structurally stable. Hence, for the topological study of nonlinear programs, it is natural to assume that the condition above holds.

In [11], under the assumption above, we obtained a necessary condition and a sufficient condition that a feasible set of a nonlinear program has a standard differential structure. In this paper we investigate a situation that the sufficient condition obtained in [11] becomes necessary and sufficient. For this purpose we study a singularity theory of feasible sets. The first and important papers on the singularity theory related to nonlinear programs are Jongen, Jonker, and Twilt [6] and Siersma [12] (1982). In [6], under a generic situation, Jongen, Jonker, and Twilt studied one-parameter families of feasible sets and developed a surgery theory of feasible sets related to the Morse theory. On the other hand, in [12], Siersma developed a singularity theory on finite determinacy, unfolding and classification for manifolds with corners, assuming that gradient vectors of active constraints are linearly independent.

We are also interested in constructing a singularity theory for nonlinear programs under the MF-condition. This construction, however, is not so easy. Hence, we have to go step by step. Therefore, in this paper we only study 1-determinacy of feasible sets and prove that this is equivalent to studying the 1-determinacy of mappings (Gibson [3]) in a restricted sense.

1.1. Notation and symbols.

R^n	n-dimensional Euclidean space,
C^∞	the class of any times continuously differentiable mappings,
$f\|U$	the restriction of a mapping f to the set U,

* Received by the editors November 21, 1988; accepted for publication (in revised form) May 1, 1990.

† National Research Institute of Fisheries Science, Kachidoki 5-5-1, Chuo-ku, Tokyo 104, Japan.

‡ Department of Management and Informatics, San-noh College, Kamikasuya 1573, Isehara, Kanagawa 259-11, Japan.

§ Department of Management Science, Science University of Tokyo, Kagurazaka 1-3, Shinjuku-ku, Tokyo 162, Japan.

$\sim_x$	the equivalent relation on the set of all C^∞-maps $\{f: R^n \to R^p\}$ at x defined by $f_1 \sim_x f_2$ if and only if there exists an open neighborhood U satisfying $f_1 \mid U = f_2 \mid U$,
D_x	differential operator with respect to x,
I_n	the identity matrix of order n,
$E_{n,p}$	the set of all C^∞ map germs of $(R^n, 0) \to R^p$,
$E^0_{n,p}$	the set of all C^∞ map germs of $(R^n, 0) \to (R^p, 0)$,
Diff_n	the set of all C^∞-diffeomorphic germs of $(R^n, 0) \to (R^n, 0)$,
Diff^1_n	the subset of $\phi \in \mathrm{Diff}_n$ with $D_x\phi(0) = I_n$,
K_n	$= \{\Psi_{\sigma,u} \in \mathrm{Diff}_n \mid$there exist a germ $u \in E_{n,n}$ such that $u_i(0) \neq 0$ $(i = 1, \cdots, n)$ and a permutation σ on $\{1, \cdots, n\}$ satisfying $\Psi_{\sigma,u}(x) = (u_1(x)x_{\sigma(1)}, \cdots, u_n(x)x_{\sigma(n)})^T\}$ where u_i is the ith component of u,
K^1_n	$= K_n \cap \mathrm{Diff}^1_n$,
$G_{n,p}$	$= \mathrm{Diff}_p \times \mathrm{Diff}_n$,
$H_{n,p}$	$= K_p \times \mathrm{Diff}_n$,
$H^1_{n,p}$	$= K^1_p \times \mathrm{Diff}^1_n$,
$f \approx_G g$	if and only if $G_{n,p} \cdot f = G_{n,p} \cdot g$ (A-equivalence),
$f \approx_H g$	if and only if $H_{n,p} \cdot f = H_{n,p} \cdot g$ (restricted A-equivalence),
$f \approx_{H^1} g$	if and only if $H^1_{n,p} \cdot f = H^1_{n,p} \cdot g$ (strongly A-equivalence),
$j^1 f$	the Taylor expansion of order one of f at $x = 0$ for $f \in E^0_{n,p}$ (the 1-jet of f),
$J^1 f$	$= \{g \in E^0_{n,p} : j^1 g = j^1 f\}$,
f	is 1-determined if and only if $J^1 f \subset G_{n,p} \cdot f$,
f	is restricted 1-determined if and only if $J^1 f \subset H_{n,p} \cdot f$,
f	is strongly restricted 1-determined if and only if $J^1 f \subset H^1_{n,p} \cdot f$,
$M[g]$	$= \{x \in R^n$: $g(x) \geqq 0$ for a C^∞ mapping g: $(R^n, 0) \to (R^p, 0)\}$,
$M[g]_0$	is the feasible set germ,
$M[\ ]_0$	is the set of all feasible set germs at $0 \in R^n$,
δ^i_j	is Kronecker's delta,
$M(p, n : R)$	is the set of real $p \times n$ matrices,
R_+	the set of nonnegative numbers,
$C_g(0)$	the cotangent cone of $M[g]_0$,
B	$= \{\xi_k\}^p_{k=1}$: a basis for $C_g(0)$,
B_k	$= \{j$: $Dg_j(0) \in R_+\xi_k\}$,
$G_k(x)$	$= \min\{g_j(x) : j \in B_k\}$,
class C	$= \{g \in E^0_{n,p}$: $M[g]_0 = M[G]_0\}$,
$C_g(0)^d$	the dual cone (the tangent cone) of $M[g]_0$.

2. Restricted 1-determinacy of C^∞-map germs. Let R^n be the n-dimensional Euclidean space. For a given point $\bar{x} \in R^n$, define an equivalence relation $\sim_{\bar{x}}$ on the set of all any times continuous differentiable (C^∞) maps $\{f: R^n \to R^p\}$ by $f_1 \sim_{\bar{x}} f_2$ if there exists an open neighborhood U of $\bar{x}$ that satisfies $f_1 | U = f_2 | U$. Here $f | U$ denotes the restriction of a mapping f to U. A C^∞-map germ from R^n to R^p at $\bar{x}$ is defined as an equivalence class with respect to $\sim_{\bar{x}}$. If f_1 and f_2 are two representatives of a germ f at $\bar{x}$, then obviously $f_1(\bar{x}) = f_2(\bar{x}) (= \bar{y})$. This same value $\bar{y}$ is called the value of the germ f at $\bar{x}$ and denoted by $f(\bar{x})$. To show this explicitly, we write $f: (R^n, \bar{x}) \to (R^p, \bar{y})$.

Let $E_{n,p}$ be the set of all C^∞-map germs at zero of $R^n \to R^p$, $E^0_{n,p}$ be the subset of $f \in E_{n,p}$ with $f(0) = 0$, Diff_n be the set of all C^∞-diffeomorphic germ ϕ's of $(R^n, 0) \to (R^n, 0)$, and Diff^1_n be the subset of Diff_n such that $D_x\phi(0) = I_n$, where D_x denotes the partial differential operator with respect to x and I_n is the unit matrix of order n.

Define a subset of Diff_n by $K_n := \{\Psi_{\sigma,u} \in \mathrm{Diff}_n \mid$ there exist a germ $u \in E_{n,n}$ such that $u_i(0) \neq 0$ $(i=1,\cdots,n)$ and a permutation σ on $\{1,\cdots,n\}$ satisfying $\Psi_{\sigma,u}(x) = (u_1(x)x_{\sigma(1)},\cdots,u_n(x)x_{\sigma(n)})^T\}$, where u_i is the ith component of u and T denotes the transpose of vectors. Moreover, define a subset of Diff_n by $K_n^1 := K_n \cap \mathrm{Diff}_n^1 = \{\Psi_{\sigma,u} \in \mathrm{Diff}_n^1 \mid u_i(0) = 1$ $(i=1,\cdots,n)$ and $\sigma = \mathrm{id}\}$ where id is the identical permutation. Let $G_{n,p} := \mathrm{Diff}_p \times \mathrm{Diff}_n$, $H_{n,p} := K_p \times \mathrm{Diff}_n$ and $H_{n,p}^1 := K_p^1 \times \mathrm{Diff}_n^1$. Then it is clear that $H_{n,p}^1 \subseteq H_{n,p} \subseteq G_{n,p}$ as subgroups. $G_{n,p}$ acts on $E_{n,p}^0$ from left and right as follows. For any $\alpha = (\beta,\gamma) \in G_{n,p}$ and $f \in E_{n,p}^0$, $\alpha \cdot f := \beta \circ f \circ \gamma^{-1}$, i.e., the diagram below is commutative:

$$\begin{tikzcd}
R^n \arrow[r, "f"] \arrow[d, "\gamma"'] & R^p \arrow[d, "\beta"] \\
R^n \arrow[r, "\alpha \cdot f"'] & R^p.
\end{tikzcd}$$

Since $H_{n,p}$ and $H_{n,p}^1$ are subgroups of $G_{n,p}$, they also act on $E_{n,p}^0$ from left and right. Suppose f and $g(\in E_{n,p}^0)$ are in the same orbit by $G_{n,p}$ (respectively, $H_{n,p}$, $H_{n,p}^1$), i.e., $G_{n,p} \cdot g = G_{n,p} \cdot f$ (respectively, $H_{n,p} \cdot g = H_{n,p} \cdot f$, $H_{n,p}^1 \cdot g = H_{n,p}^1 \cdot f$); then we write $f \approx_G g$ (respectively, $f \approx_H g$, $f \approx_{H^1} g$). Since it is obvious that these relations are equivalence relations, we call these equivalence relations A-equivalence (see, for instance, Gibson [3]), restricted A-equivalence, and strongly restricted A-equivalence, respectively. We will mainly investigate the (strongly) restricted A-equivalence in the following because it takes an important role in studying the singularities of feasible sets in nonlinear programs.

Denote the 1-jet of $f \in E_{n,p}^0$ at $0 \in R^n$ by $j^1 f := (D_x f_1(0)x,\cdots,D_x f_p(0)x)$ and put $J_f^1 := \{g \in E_{n,p}^0 \mid j^1 g = j^1 f\}$. If $J_f^1 \subseteq G_{n,p} \cdot f$ (respectively, $H_{n,p} \cdot f$, $H_{n,p}^1 \cdot f$) holds, then f is called 1-determined (respectively, restricted 1-determined, strongly restricted 1-determined).

For a $g \in E_{n,p}^0$, denote $\{x \in R^n \mid g_i(x) \geqq 0,\ i=1,2,\cdots,p\}$ by $M[g]$. $M[g]$ is called the feasible set defined by g. For $0 \in R^n$, define an equivalence relation on the set $\{M[g] \mid g \in E_{n,p}^0\}$ by $M[g_1] \sim M[g_2]$ if there exists a neighborhood U of $0 \in R^n$ such that $M[g_1] \cap U = M[g_2] \cap U$. A feasible set germ on R^n at $0 \in R^n$ is defined as an equivalence class with respect to $\sim$ and denoted by $M[g]_0$. Denote the set of all feasible set germs at $0 \in R^n$ by $M[\ ]_0$. It is clear that for a germ $g \in E_{n,p}^0$ and any representatives g_1 and g_2 of g, $M[g_1] \sim M[g_2]$. Hence we may define a feasible set germ $M[g]_0$ by $M[g_1]_0$ for any C^∞-map germ $g \in E_{n,p}^0$ and its representative g_1. Let ρ be a map assigning $g \in E_{n,p}^0$ to $M[g]_0$. It is also clear that ρ is onto from $E_{n,p}^0$ to $M[\ ]_0$. We say that $M[g]_0$ is 1-determined if $M[g]_0$ is diffeomorphic to $M[h]_0$ by the action of Diff_n (denoted by $M[g]_0 \cong M[h]_0$) whenever $M[j^1 g]_0 = M[j^1 h]_0$.

Remark 2.1. Note that ρ is not one to one. For example, let $g, h \in E_{n,p}^0$ be given by $g_i = \delta_i^1 x_1$ and $h_i = \delta_i^1 x_1^3$ for $i = 1,2,\cdots,p$, where $\delta_i^j = 1$ $(i=j)$, or 0 (otherwise). Then $g \not\approx h$ but $M[g]_0 = M[h]_0$.

With respect to 1-determinancy of C^∞-map germs, the lemma below holds.

LEMMA 2.2. *Let $f \in E_{n,p}^0$. Then f is restricted* 1*-determined if and only if f is strongly restricted* 1*-determined.*

Proof. It is sufficient to show the only if part, since the if part is clear. Suppose $J_f^1 \subset H_{n,p} \cdot f$. Let $g \in J_f^1$. Note that $D_x g(0)x = D_x f(0)x \in J_f^1$. Then, by the assumption, there exist $\alpha_i = (\beta_i, \phi_i) \in H_{n,p}$ $(i=1,2)$ such that

$$\beta_1 \circ f \circ \phi_1^{-1}(x) = D_x f(0)x = D_x g(0)x = \beta_2 \circ g \circ \phi_2^{-1}(x),$$

i.e.,

$$\beta_1 \circ f(x) = \langle D_x f(0), \phi_1(x) \rangle$$

and

$$\beta_2 \circ g(x) = \langle D_x g(0), \phi_2(x) \rangle.$$

Let $\beta_1(y) = (u_1(y) y_{\sigma(1)}, \cdots, u_p(y) y_{\sigma(p)})$, where $u_i(0) \neq 0$ $(1 \leqq i \leqq p)$ and σ is a permutation on $\{1, \cdots, p\}$. By differentiating $u_i(f(x)) f_{\sigma(i)}(x) = \langle D_x f_i(0), \phi_1(x) \rangle$, we have $u_i(0) D_x f_{\sigma(i)}(0) = D_x f_i(0) D_x \phi_1(0)$. Hence $D_x f_i(0) = u_i(0) D_x f_{\sigma(i)}(0)(D_x \phi_1(0))^{-1}$. This leads to

$$(u_i(f(x))/u_i(0)) f_{\sigma(i)}(x) = \langle D_x f_{\sigma(i)}(0), (D_x \phi_1(0))^{-1} \phi_1(x) \rangle.$$

Set $\tau = \sigma^{-1}$, $\tilde{u}_i(x) = u_{\tau(i)}(x)/u_{\tau(i)}(0)$ and $\tilde{\phi}(x) = (D_x \phi_1(0))^{-1} \phi_1(x)$; then we get $\tilde{u}_i(f(x)) f_i(x) = \langle D_x f_i(0), \tilde{\phi}(x) \rangle$. Obviously, $\tilde{u}_i(0) = 1$ and $D_x \tilde{\phi}(0) = I_n$. By repeating the same argument for $D_x g(0) x = \beta_2 \circ g \circ \phi_2^{-1}(x)$, we see that there exist $\tilde{\alpha}_i = (\tilde{\beta}_i, \tilde{\phi}_i) \in H^1_{n,p}$ $(i = 1, 2)$ satisfying

$$\tilde{\beta}_1 \circ f \circ \tilde{\phi}_1^{-1}(x) = D_x f(0) x = D_x g(0) x = \tilde{\beta}_2 \circ g \circ \tilde{\phi}_2^{-1}(x).$$

Define $\bar{\beta}(y) = \tilde{\beta}_2^{-1} \circ \tilde{\beta}_1(y)$ and $\bar{\phi}(x) = \tilde{\phi}_2^{-1} \circ \tilde{\phi}_1(x)$. Then $g(x) = \bar{\beta} \circ f \circ \bar{\phi}^{-1}(x)$. It is easily shown that $(\bar{\beta}, \bar{\phi}) \in H^1_{n,p}$. □

The following four definitions are fundamental in what follows. Let $M(p, n : R)$ be the set of real $p \times n$ matrices.

DEFINITION 2.3. Let $A \in M(p, n : R)$. We say that A is of type 1 if

(1) rank $A = \min(p, n)$,

(2) $p \leqq n + 2$,

(3) All the minors of degree n of A are nonzero.

DEFINITION 2.4. Let $A \in M(p, n : R)$. We say A is of type (*) if for any $g \in E^0_{n,p}$ with $D_x g(0) = A$; then there exist $z \in E_{n,p}$ such that $z_i(0) \neq 0$ $(1 \leqq i \leqq p)$ and $\phi \in \mathrm{Diff}_n$ satisfying $z_i(x) g_i(x) = \langle D_x g_i(0), \phi(x) \rangle$ $(1 \leqq i \leqq p)$.

DEFINITION 2.5. Let $g \in E^0_{n,p}$. We say that g is irredundant if $D_x g_i(0)$ and $D_x g_j(0)$ are linearly independent to each other for $i \neq j$.

Remark 2.6. The definition of "irredundant" is related to that of "smoothness condition" in [11].

DEFINITION 2.7. Let $g \in E^0_{n,p}$. We say that g is a minimal representative if $M[g]_0 \neq M[g^{\langle i \rangle}]_0$ for all i where $g^{\langle i \rangle}$ is defined by $g_j^{\langle i \rangle} = (1 - \delta^i_j) g_j$.

Remark 2.8. The definition of "a minimal representative" is related to that of "a conical basis" in [11].

When we study topological properties of $M[g]$, it is essential to assume that the well-known MF-condition holds (see, for example, Guddat and Jongen [4], Guddat, Jongen, and Rueckmann [5]):

Condition 2.9. Let $A \in M(p, n : R)$. Then there exists a $w \in R^n$ such that $Aw > 0$ holds.

Condition 2.10 (Mangasarian and Fromovitz [10]). Let $g \in E^0_{n,p}$. Then $D_x g(0)$ satisfies Condition 2.9.

Remark 2.11. Since we do not treat equality constraints, the MF-condition coincides with the Cottle constraints qualification (Gauvin [2]).

According to [11], we restrict the problem to the class below. Let $g \in E^0_{n,p}$. Define the cotangent cone $C_g(0)$ by

$$C_g(0) = \left\{ \sum_{1 \leqq i \leqq p} r_i D_x g_i(0) : r_i \geqq 0 \ (1 \leqq i \leqq p) \right\}.$$

Since $C_g(0)$ is finitely generated, there is a nonnegative integer q such that a basis of $C_g(0)$ is $B := \{\xi_k\}_{k=1}^q$. It is easily shown that q is independent of the choice of a basis. Define the index subset by $B_k := \{j: D_x g_j(0) \in R_+ \xi_k\}$ and the minimal constraint function by $G_k(x) := \min\{g_j(x): j \in B_k\}$, where $R_+ = \{r \in R: r \geqq 0\}$. $M[g]_0 = M[G]_0$ does not necessarily hold. So, in the remainder of the paper, we assume that $g \in E_{n,p}^0$ is always in the class $C = \{g \in E_{n,p}^0: g$ satisfies $M[g]_0 = M[G]_0\}$.

We will introduce two main results of the paper [11] in the following theorem.

THEOREM 2.12 [11]. *Let $n \geqq 2$ and $A \in M(p, n: R)$. Suppose the matrix A satisfies Condition* 2.9; *then A is of type* 1 *if and only if A is of type* (*).

Proof. Only if. Let A be of type 1. If $p \leqq n$, then rank $D_x g(0) = p$, where $g \in E_{n,p}^0$ with $D_x g(0) = A$. Hence it is easy to show the existence of $z \in E_{n,p}$ with $z_i(0) \neq 0$ $(1 \leqq i \leqq p)$ and $\phi \in \mathrm{Diff}_n$ such that $z_i(x) g_i(x) = \langle D_x g_i(0), \phi(x)\rangle$ $(1 \leqq i \leqq p)$. (See, for example, Jongen, Jonker, and Twilt [7].) If $p > n$, i.e., $p = n+1$ or $n+2$, the proof is constructive. In fact, we can construct $z(x)$ and $\phi(x)$ for any $g \in E_{n,p}^0$ with $D_x g(0) = A$ in the category of C^∞ map germs. For more details, see Theorem 4.2 of [11].

If. Suppose that A is of type (*). It is easily shown that rank $A = \min(p, n)$. If $p \leqq n$, it is clear that A is of type 1. If $p > n$, we must treat this carefully. In this case, without loss of generality, we may assume that A is of the form $[I_n, B^T]^T$, where B is a matrix of an appropriate size. By constructing $g \in E_{n,p}^0$ such that $D_x g(0) = A$ and each g_i $(1 \leqq i \leqq p)$ is a linear or a quadratic map germ, we can verify that each entry of B is nonzero, rank $B \leqq 2$ and all the minors of B with order two are not equal to zero. If $p \geqq n+3$, then it is easy to construct a g with $D_x g(0) = A$, leading to a contradiction. For more details, see Lemma 3.4 and Theorem 3.2 of [11]. □

THEOREM 2.13 (characterization of 1-determinacy of feasible sets). *Let $n \geqq 2$ and $g \in E_{n,p}^0$ be irredundant, a minimal representative, and satisfy Condition* 2.10. *Then $M[g]_0$ is* 1*-determined if and only if $D_x g(0)$ is of type* 1.

Proof. Only if. Suppose that $M[g]_0$ is 1-determined. Then $M[g]_0 \cong M[j^1 g]_0$. By Theorem 3.1 of [11], it follows that there exist $z_i(x)$ $(1 \leqq i \leqq p)$ satisfying

$$g_i(x) z_i(x) = \langle D_x g_i(0), y(x)\rangle \qquad (1 \leqq i \leqq p)$$

where $y(x)$ is a local diffeomorphism from $M[g]_0$ to $M[j^1 g]_0$ and $y(0) = 0$. This implies that $D_x g(0)$ is of type (*). By Theorem 2.12, we see that $D_x g(0)$ is of type 1.

If. Suppose that $D_x g(0)$ is of type 1. Define the cotangent cone $C_g(0)$ and its dual cone (the tangent cone) $C_g(0)^d$ of $M[g]_0$ by

$$C_g(0) = \left\{ \sum_{1 \leqq i \leqq p} r_i D_x g_i(0) : r_i \geqq 0 \ (1 \leqq i \leqq p) \right\}$$

and

$$C_g(0)^d = \{v \in R^n : \langle v, u\rangle \geqq 0 \text{ for every } u \in C_g(0)\}.$$

Note that $M[j^1 g]_0 = M[j^1 h]_0$ means $C_g(0)^d = C_h(0)^d$). (See Appendix D in [11].)

Let $p \leqq n$. Then rank $D_x g(0) = p$. In this case, there exists a local diffeomorphism y_g with $y_g(0) = 0$ from $M[g]_0$ to $C_g(0)^d$. (See, for example, Jongen, Jonker, and Twilt [7].) From the fact that $g(0) = h(0) = 0 \in R^p$ and rank $D_x g(0) = p$, we see that rank $D_x h(0) = p$. Hence there exists a local diffeomorphism y_h from $M[h]_0$ to $C_h(0)^d$. Since $C_g(0)^d = C_h(0)^d$, it follows that $y_h^{-1} \circ y_g$ is a local diffeomorphism from $M[g]_0$ to $M[h]_0$, i.e., $M[g]_0$ is 1-determined.

Next, suppose $p \geqq n$. By Theorem 4.2 of [11], we can construct a local diffeomorphism y_g from $M[g]_0$ to $C_g(0)^d$ since $D_x g(0)$ is of type 1. The assumptions with respect to g and $M[j^1 g]_0 = M[j^1 h]_0$ imply rank $D_x h(0) = n$. It is easily shown that $D_x h(0)$ is also of type 1. Repeating the same argument above, we obtain a local diffeomorphism y_h of $M[h]_0$ to $C_h(0)^d$. Thus $y_h^{-1} \circ y_g$ is a local diffeomorphism from $M[g]_0$ to $M[h]_0$ since $C_g(0)^d = C_h(0)^d$. Therefore $M[g]_0$ is 1-determined. □

In order to get an understanding of Theorem 2.13, we give two examples with $g:(R^3, 0) \to (R^5, 0)$ below.

Example 2.14. Let $g_i(x) = x_i$ $(1 \leqq i \leqq 3)$, $g_4(x) = 2x_1 + x_2 - x_3$, and $g_5(x) = x_1 + x_1^2 + 2x_2 - x_3$. Then $M[g]_0$ is 1-determined. (Note that $D_xg(0)$ is of type 1.) To show this, we must construct a $\phi \in \mathrm{Diff}_3$ and $z_i(x)$ $(1 \leqq i \leqq 5)$ with $z_i(0) \neq 0$ satisfying

$$z_i(x)g_i(x) = \langle D_xg_i(0), \phi(x)\rangle \qquad (1 \leqq i \leqq 5).$$

Assuming $z_3(x) = 1$ with no loss of generality, a tedious calculation (see Appendix B in [11]) shows that

$$z_1(x) = (3 + 3x_1 + 2x_2 - 3x_3)/3(1 - x_1),$$
$$z_2(x) = (3 - x_1)/3(1 - x_1), \qquad z_3(x) = 1,$$
$$z_4(x) = (1 + x_1)/(1 - x_1), \qquad z_5(x) = 1/(1 - x_1)$$

and $\phi = (\phi_1, \phi_2, \phi_3)$ is given by

$$\phi_1(x) = x_1z_1(x), \quad \phi_2(x) = x_2z_2(x), \quad \phi_3(x) = x_3z_3(x).$$

Example 2.15. Let $g_i(x) = x_i$ $(1 \leqq i \leqq 3)$, $g_4(x) = 2x_1 + x_2 - x_3$, and $g_5(x) = x_1 + x_2 - x_3 - x_3^2$. Then $M[g]_0$ is not 1-determined. (Note that $D_xg(0)$ is not of type 1.) Suppose there exists a $\phi \in \mathrm{Diff}_3$ and $z_i(x)$ $(1 \leqq i \leqq 5)$ such that

$$z_i(x)g_i(x) = \langle D_xg_i(0), \phi(x)\rangle \qquad (1 \leqq i \leqq 5).$$

Then

$$\phi_1(x) = x_1z_1(x), \quad \phi_2(x) = x_2z_2(x), \quad \phi_3(x) = x_3z_3(x),$$
$$2\phi_1(x) + \phi_2(x) - \phi_3(x) = (2x_1 + x_2 - x_3)z_4(x),$$
$$\phi_1(x) + \phi_2(x) - \phi_3(x) = (x_1 + x_2 - x_3 - x_3^2)z_5(x).$$

Without loss of generality, we may assume $z_3(x) = 1$. From equations above, we obtain

$$(*) \qquad 2x_1z_1(x) + x_2z_2(x) - x_3 = (2x_1 + x_2 - x_3)z_4(x)$$

and

$$(**) \qquad x_1z_1(x) + x_2z_2(x) - x_3 = (x_1 + x_2 - x_3 - x_3^2)z_5(x).$$

Consider the Taylor expansion at 0 of each $z_i(x)$. Put

$$z_1(x) = a_1 + s_1x_1 + s_2x_2 + s_3x_3 + \cdots,$$
$$z_2(x) = a_2 + t_1x_1 + t_2x_2 + t_3x_3 + \cdots,$$
$$z_4(x) = a_4 + u_1x_1 + u_2x_2 + u_3x_3 + \cdots,$$
$$z_5(x) = a_5 + v_1x_1 + v_2x_2 + v_3x_3 + \cdots.$$

Substitute these Taylor expansions into (*) and (**). Then it is easily shown that $a_1 = a_2 = a_4 = a_5 = 1$. Furthermore, we have $t_2 + t_3 = 0$ from (*) and $t_2 + t_3 + 1 = 0$ from (**). This is a contradiction.

Now we are in a position to characterize the restricted 1-determinacy of C^∞-map germs.

LEMMA 2.16. *Let $f \in E^0_{n,p}$. Suppose f is strongly restricted* 1*-determined*; *then $D_xf(0)$ is of type* (*).

Proof. We must prove that $A = D_xf(0)$ has the property (*) of Definition 2.3. Suppose that $g \in E_{n,p}^0$ satisfies $D_xg(0) = A(= D_xf(0))$. Since f is strongly restricted 1-determined and $D_xg(0) = D_xf(0)$, g is also strongly restricted 1-determined. This implies $g(x) \approx_{H^1} D_xg(0)x$. Hence there exist $\phi \in \mathrm{Diff}_n^1$ and $u \in E_{n,p}$ such that $u_i(0) = 1$ and $g_i(x)u_i(g(x)) = \langle D_xg_i(0), \phi(x)\rangle$ for $i = 1, 2, \cdots, p$. Set $z_i(x) := u_i(g(x))$ for $1 \leqq i \leqq p$ and the lemma holds. □

THEOREM 2.17. *Let $n \geqq 2$ and $f \in E_{n,p}^0$. Suppose f satisfies Condition* 2.10; *then the following two conditions are equivalent*:

(1) *f is restricted* 1*-determined,*

(2) *$D_xf(0)$ is of type* 1.

Proof. From Theorem 2.12 and Lemmas 2.4 and 2.16, (1) clearly implies (2). Conversely, we will consider the case that condition (2) holds. Suppose that $g \in E_{n,p}^0$ satisfies $D_xg(0) = D_xf(0)$. If $p < n$, then $\operatorname{rank} D_xg(0) = \operatorname{rank} D_xf(0) = \min(p, n) = p$. Hence there exist $\tilde{f}, \tilde{g} \in \mathrm{Diff}_n$ such that $f = \nu \circ \tilde{f}$ and $g = \nu \circ \tilde{g}$, where ν is the natural projection from R^n to R^p. Define $\phi \in \mathrm{Diff}_n$ by $\tilde{g}^{-1} \circ \tilde{f}$. Then $\tilde{f} = \tilde{g} \circ \phi$. Hence $f = \nu \circ \tilde{f} = \nu \circ \tilde{g} \circ \phi = g \circ \phi$. This shows that $f \approx_H g$. When $p \geqq n$, from Theorem 2.12, there exist $z \in E_{n,p}$ such that $z_i(0) \neq 0$ and $\phi \in \mathrm{Diff}_n$ such that $g_i(x)z_i(x) = \langle D_xg_i(0), \phi(x)\rangle$ for $1 \leqq i \leqq p$. Since $p \geqq n$, $\operatorname{rank} D_xg(0) = n$. Hence the equation $y = g(x)$ can be solved with respect to x, i.e., there exists $h \in E_{p,n}^0$ such that locally $h \circ g = \mathrm{Id}_{R^n}$. Therefore $g_i(x)z_i(h(g(x))) = \langle D_xg_i(0), \phi(x)\rangle$ for $1 \leqq i \leqq p$. This means $g \approx_H j^1g(0) = D_xg(0)x$. Similarly, we have $f \approx_H j^1f(0) = D_xf(0)x$. Since $D_xg(0) = D_xf(0)$, we obtain $g \approx_H j^1g(0) = j^1f(0) \approx_H f$. □

COROLLARY 2.18. *Let $n \geqq 2$ and $g \in E_{n,p}^0$ be irredundant, a minimal representative, and satisfy Condition* 2.10. *Then the next three statements are equivalent*:

(1) *g is restricted* 1*-determined,*

(2) *g is strongly restricted* 1*-determined,*

(3) *$M[g]_0$ is* 1*-determined.*

Proof. The proof is clear from Lemma 2.4, and Theorems 2.13 and 2.17. □

Acknowledgments. The authors thank the referees for valuable comments and suggestions.

REFERENCES

[1] O. FUJIWARA, *Morse programs: a topological approach to smooth constrained optimization.* I, Math. Oper. Res., 7 (1982), pp. 602–616.

[2] J. GAUVIN, *A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming*, Math. Programming, 12 (1977), pp. 136–138.

[3] C. G. GIBSON, *Singular Points of Smooth Mappings*, Research Notes in Mathematics, Vol. 25, Pitman, Boston, 1979.

[4] J. GUDDAT AND H. TH. JONGEN, *Structural stability in nonlinear optimization*, Optimization, 18 (1987), pp. 617–631.

[5] J. GUDDAT, H. TH. JONGEN, AND J. RUECKMANN, *On stability and stationary points in nonlinear optimization*, J. Australian Math. Soc. Ser. B, (1986), pp. 36–56.

[6] H. TH. JONGEN, P. JONKER, AND F. TWILT, *On one-parameter families of sets defined by (in)equality constraints*, Nieuw Arch. Wisk. (3), 30 (1982), pp. 307–322.

[7] ———, *Nonlinear Optimization in R^n*, I. *Morse Theory, Chebyshev Approximation*, Peter Lang Verlag, Frankfurt a.M., Bern, New York, 1983.

[8] ———, *Nonlinear Optimization in R^n*, II. *Transversality, Flows, Parametric Aspects*, Peter Lang Verlag, Frankfurt a.M., Bern, New York, 1986.

[9] M. KOJIMA, *Strongly stable stationary solutions in nonlinear programs*, in Analysis and Computation of Fixed Points, S. M. Robinson, ed., Academic Press, New York, 1980.

[10] O. L. MANGASARIAN AND S. FROMOVITZ, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl., 17 (1967), pp. 37-47.

[11] T. MATSUMOTO, S. SHINDOH, AND R. HIRABAYASHI, *Local Linearization of Feasible Sets*, Research Reports on Information Sciences, B-210, Tokyo Institute of Technology, Tokyo, Japan, 1988.

[12] D. SIERSMA, *Singularities of functions on boundaries, corners, etc.*, Quart. J. Math., Oxford Ser. (2), 32 (1982), pp. 119-127.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 347-367, March 1991

006

ASYMPTOTIC LOCATIONS OF EIGENFREQUENCIES OF EULER–BERNOULLI BEAM WITH NONHOMOGENEOUS STRUCTURAL AND VISCOUS DAMPING COEFFICIENTS*

HANKUN WANG† AND GOONG CHEN‡

Abstract. A slender beam has two spatially nonhomogeneous damping terms. The first one acts opposite to the bending moment time derivative and is sometimes called *structural damping*, while the second acts opposite to the velocity and is called *viscous damping*. When these damping coefficients are constant, it is known that structural damping causes a strong attenuation rate that is frequency-proportional, whereas viscous damping causes a constant attenuation rate for all frequencies. In this paper, using the method of Birkhoff [*Trans. Amer. Math. Soc.*, 9 (1908), pp. 219-231], [*Trans. Amer. Math. Soc.*, 9 (1908), pp. 373-395], and Birkhoff and Langer [*Proc. Amer. Acad. Arts Sci.*, (2) 58 (1923), pp. 51-128] explicit asymptotic expressions for the eigenfrequencies of the nonhomogeneous damping problem are derived. It is shown that the asymptotic patterns of the eigenspectrum remain similar to the constant coefficients case. The viscous damping effect is also shown to cause a constant shift to both the attenuation rates and the frequencies; thus it is overwhelmed by the structural damping effect.

Because experimentally it has been observed that all eigenfrequencies of light beams essentially lie within the asymptotic regime, the asymptotic formulas derived herein should be useful in determining the pole assignment for feedback stabilization.

Key words. beam equations, damping, locations of eigenfrequencies

AMS(MOS) subject classifications. 93D99, 93D15, 35B35, 34E05

1. Introduction.

The analysis of damping is important in the understanding of mechanical behavior and control of vibrating systems. For lightweight flexible vibrating structures in contemporary large space technology, the EB (Euler-Bernoulli) beam equation is the most commonly used mathematical model. Several papers [4], [5], [10], [11], [12] have addressed questions in the modeling and analysis of distributed damping on an EB beam, for which it has been noted that the rate of attenuation of eigenmodes is roughly proportional to the frequency. Consequently, a "square root" operator was incorporated in the EB equation to model the distributed damping effect. The simplest such modeling equation is

$$(1.1)\quad m\frac{\partial^2}{\partial t^2}y(x,t)-2m\gamma_1\frac{\partial^2}{\partial t\partial x^2}y(x,t)+EI\frac{\partial^4}{\partial x^4}y(x,t)=0,\qquad 0<x<\pi,\quad t>0,$$

where m, EI, and γ_1 are positive constants signifying, respectively, the mass density, flexural rigidity, and damping coefficient. The beam length has been chosen to be π just for the sake of some notational convenience later. Associated with (1.1) the hinged boundary conditions are used at both ends:

$$(1.2)\qquad y(0,t)=y_{xx}(0,t)=y(\pi,t)=y_{xx}(\pi,t)=0.$$

Two initial conditions $y(x,0)$, $y_t(x,0)$ will also need to be prescribed, but they will not be used here.

Let λ be an eigenfrequency and $\phi(x)$ be an eigenmode. Then

$$(1.3)\qquad y(x,t)=e^{\lambda t}\phi(x),\qquad 0\leqq x\leqq\pi,$$

* Received by the editors April 10, 1989; accepted for publication (in revised form) April 27, 1990.

† Department of Mathematics, Wichita State University, Wichita, Kansas 67208.

‡ Department of Mathematics, Texas A&M University, College Station, Texas 77840. This research was supported in part by Air Force Office of Scientific Research grant 87-0334.

leads to the following eigenvalue problem:

$$\begin{aligned}(1.4)\qquad & a^4\phi^{(4)}(x)-2\gamma_1\lambda\phi''(x)+\lambda^2\phi(x)=0, \qquad 0<x<\pi,\ a\equiv\left(\frac{EI}{m}\right)^{1/4},\\ & \phi(0)=\phi''(0)=\phi(\pi)=\phi''(\pi)=0.\end{aligned}$$

We easily derive that

$$\begin{aligned}(1.5)\qquad & \lambda=-\gamma_1k^2\pm ik^2\sqrt{a^4-\gamma_1^2}, \\ & \phi(x)=\sin kx, \qquad k=1,2,3,\cdots,\end{aligned}$$

where $k^2(a^4-\gamma_1^2)^{1/2}$ is the frequency. The spectral pattern is shown in Fig. 1. All eigenfrequencies lie on the rays which form angles

$$(1.6)\qquad \theta=\pm\tan^{-1}\frac{\gamma_1}{\sqrt{a^4-\gamma_1^2}}$$

with respect to the imaginary axis. Note that generally the value of γ is not large, thus we can assume

$$(1.7)\qquad a^4-\gamma_1^2>0.$$

More recently, Russell has proposed a more elaborate damping model with governing integrodifferential equation

$$\begin{aligned}(1.8)\qquad & m\frac{\partial^2}{\partial t^2}y(x,t)-2\frac{\partial}{\partial x}\int_0^\pi h(x,\xi)\left[\frac{\partial^2}{\partial t\partial x}y(x,t)-\frac{\partial^2}{\partial t\partial \xi}y(\xi,t)\right]d\xi \\ & \qquad +EI\frac{\partial^4}{\partial x^4}y(x,t)=0, \qquad 0<x<\pi, \quad t>0,\end{aligned}$$

to model frequency-proportional damping. The kernel h herein satisfies

$$h(x,\xi)=h(\xi,x) \quad \text{and} \quad \int_0^\pi\int_0^\pi h(x,\xi)[f(x)-f(\xi)]^2\,dx\,d\xi\geqq 0$$

for any function $f\in L^2(0,\pi)$. This model takes into account the presence of reinforced fibers and seems particularly suitable to model modern matrix materials. As of this writing, only small progress has been made in analyzing the asymptotic location of

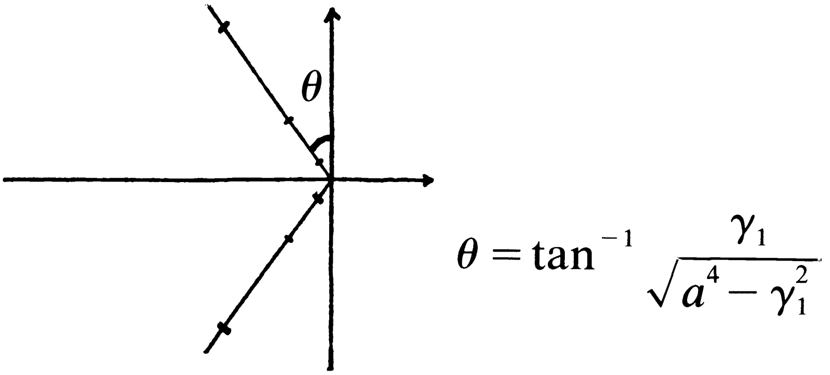


FIG. 1. *Spectral pattern of frequency proportional damping.*

the eigenspectrum, or the holomorphic semigroup property for the general equation (1.6).

We can regard (1.8) as an EB equation *with essentially a variable damping coefficient* for the term $(\partial^3/\partial t\partial x^2)y(x, t)$, with the understanding that such a coefficient is now a convoluted integrodifferential operator. Therefore to analyze the eigenspectrum pattern of (1.8) it is fundamental that we understand the corresponding problems for a simpler damped EB equation

$$m\frac{\partial^2}{\partial t^2}y(x, t)-2m\gamma_1(x)\frac{\partial^3}{\partial t\partial x^2}y(x, t)+EI\frac{\partial^4}{\partial x^4}y(x, t)=0, \tag{1.9}$$

where $\gamma_1(x)\geqq 0$ is now a *nonhomogeneous damping coefficient.* As a matter of fact, using the basic asymptotic analysis developed herein and by using this idea of viewing the convoluted integrodifferential term in Russell's equation (1.8) as something like a spatial varying damping coefficient in equation (1.9), in a recent conference proceedings [13] Wang has been able to derive an *identical* asymptotic formula (cf. (3.23)) for Russell's model (1.8) subject to the hinged boundary conditions (1.2). From the purely mathematical point of view, the eigenspectrum pattern of this equation is worth investigation in its own right as (1.9) is now much more general than (1.1).

From the control theoretic point of view, in the feedback stabilization of a vibrating structure, it is often necessary to study closed-loop systems with variable coefficients. For example, consider a controlled beam

$$m\frac{\partial^2}{\partial t^2}y(x, t)+EI\frac{\partial^4}{\partial x^4}y(x, t)=\gamma_1(x)u(x, t), \tag{1.10}$$

where $\gamma_1(x)\geqq 0$ is a locally supported function (i.e., the support of γ_1 is a proper subset of $(0, \pi)$), signifying the span on the beam interval $(0, \pi)$ where the controller is distributed (cf. Fig. 2).

If we feedback the rate of bending moment

$$u(x, t)=2m\frac{\partial^3}{\partial t\partial x^2}y(x, t)$$

in (1.10), then the closed-loop system will be just (1.9). The classical separation of variables approach will not give us the *precise* locations of the spectrum because equation (1.9) contains variable coefficients and does not have closed form solutions in general. Therefore the best hope lies in asymptotic solutions. Although the expressions we have derived herein are *asymptotic formulas,* there is both theoretical and experimental evidence from our work elsewhere [6], [7], [9] on beam theory that the eigenspectrum will be *in the asymptotic regime even for very low frequencies.* Therefore the analysis of locations of eigenfrequencies and their damping rates carried out in this paper may provide vital information for the *pole assignment* of the closed-loop

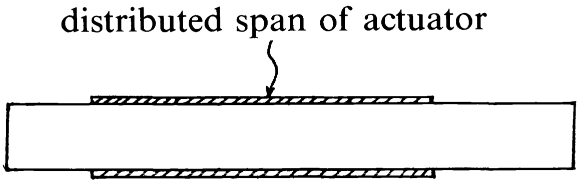


FIG. 2. *A beam with controller distributed on only part of the span.*

system. In this paper, we will actually move one step further by studying the model

$$\text{(1.11)}\qquad m\frac{\partial^2}{\partial t^2}y(x,t)-2m\gamma_1(x)\frac{\partial^3}{\partial t\partial x^2}y(x,t)+2m\gamma_2(x)\frac{\partial}{\partial t}y(x,t)+EI\frac{\partial^4}{\partial x^4}y(x,t)=0,$$

$$\gamma_1(x)\geqq 0,\quad \gamma_2(x)\geqq 0,\quad 0\leqq x\leqq \pi,\ t>0,$$

which takes into account two types of variable coefficient damping forces, the first one, $-2m\gamma_1(x)\partial^3 y/\partial t\partial x^2$, is proportional to the time derivative of bending and is sometimes termed *structural damping* (cf. [5], [11]), while the second one, $2m\gamma_2(x)\partial y/\partial t$, is proportional to velocity and is commonly referred to as *viscous damping*. As $\gamma_1(x)$ is not large, let us further assume

$$\text{(1.12)}\qquad a^4-\gamma_1^2(x)\geqq\delta>0\quad\text{on } [0,\pi],\quad \text{for some } \delta>0.$$

Our treatment and results here are also good for the equation

(1.11)′

$$m\frac{\partial^2}{\partial t^2}y(x,t)-2m\frac{\partial}{\partial x}\left[\gamma_1(x)\frac{\partial^2}{\partial t\,\partial x}y(x,t)\right]+2m\gamma_2(x)\frac{\partial}{\partial t}y(x,t)+EI\frac{\partial^4}{\partial x^4}y(x,t)=0,$$

which is perhaps an even more natural analogue of (1.8). See Remark 3.5.

If in (1.11), we let $\gamma_1(x)\equiv 0$ and $\gamma_2(x)\equiv\gamma_2>0$ on $[0,\pi]$:

$$\text{(1.13)}\qquad m\frac{\partial^2}{\partial t^2}y(x,t)+2m\gamma_2\frac{\partial}{\partial t}y(x,t)+EI\frac{\partial^4}{\partial x^4}y(x,t)=0$$

subject to the same boundary conditions (1.2), then a simple separation of variables argument shows that all eigenfrequencies λ satisfy

$$\text{(1.14)}\qquad \lambda=-\gamma_2\pm i\sqrt{a^4k^4-\gamma_2^2},\qquad k\in\mathbb{Z}.$$

Thus all sufficiently high eigenfrequencies have a uniform stability margin $-\gamma_2$.

Our paper is also motivated by an earlier study in Chen et al. [8], which states that for the second-order wave equation with a nonhomogeneous viscous damping coefficient:

$$\text{(1.15)}\qquad \frac{\partial^2 w(x,t)}{\partial t^2}+2\gamma(x)\frac{\partial w(x,t)}{\partial t}-c^2\frac{\partial^2 w(x,t)}{\partial x^2}=0,\qquad 0<x<\pi,\ t>0,$$

subject to various conservative boundary conditions at 0 and π,

the rate of damping will be the average at high frequencies

$$\text{(1.16)}\qquad -\frac{1}{\pi}\int_0^\pi \gamma(x)\,dx,$$

thus asymptotically the rate of damping caused by the variable coefficient $\gamma(x)$ in (1.15) is "homogenized" to become its average. For our system (1.11), (1.2) under study, we wonder what kind of "homogenization" result will hold for its eigenspectrum. To be precise, we pose the following questions:

(Q1) Consider equation (1.9) subject to (1.2). For large eigenfrequencies $\lambda=\mu+i\nu$, $\mu,\nu\in\mathbb{R}$, will the slope asymptotically satisfy (a homogenization property like (1.6))

$$\lim_{|\lambda|\to\infty}\frac{\mu}{\nu}=\pm\tan^{-1}\frac{\gamma_{1,av}}{\sqrt{a^4-\gamma_{1,av}^2}},$$

where

$$\gamma_{1,av} \equiv \frac{1}{\pi}\int_0^\pi \gamma_1(x)\,dx?$$

(Q2) Consider the equation

$$m\frac{\partial^2}{\partial t^2}y(x,t)+2m\gamma_2(x)\frac{\partial}{\partial t}y(x,t)+EI\frac{\partial^4}{\partial x^4}y(x,t)=0 \tag{1.17}$$

subject to boundary conditions (1.2). For large eigenfrequencies $\lambda = \mu+i\nu$, μ, $\nu \in \mathbb{R}$, will the stability margin satisfy (a homogenization property like (1.16))

$$\lim_{|\lambda|\to\infty} \mu = -\gamma_{2,av},$$

where

$$\gamma_{2,av} \equiv \frac{1}{\pi}\int_0^\pi \gamma_2(x)\,dx?$$

An answer to (Q2) was also briefly mentioned in [8] without any rigorous treatment.

The main objective of this paper is to determine the asymptotic locations of the eigenfrequencies of (1.11), and to answer (Q1) and (Q2) in order to apply to various modeling, stabilization and control problems mentioned earlier. Our contributions are the following:

(i) We have shown that at large frequencies a spectral pattern like Fig. 1 will emerge. Asymptotically, all large eigenvalues will fall in a close vicinity of two rays on the left halfplane whose angle (i.e., θ, cf. Fig. 1) is determinable by $\gamma_1(x)$.

(ii) We are able to compare the magnitudes of contributions by $\gamma_1(x)$ and $\gamma_2(x)$ in the asymptotic estimates of eigenfrequencies λ.

The idea for our paper came from the work [8] cited earlier based on an asymptotic expansion procedure of Birkhoff [1], [2] and Birkhoff and Langer [3]. The eigenvalue problem treated here constitutes a fourth-order problem, but its complexity and difficulty seem to have quadrupled (instead of being merely doubled). The form of the asymptotic solution also looks rather different from the one in [8] because equation (1.1) is not homogeneous of the same degree in x and t. An additional transformation (cf. (2.4) in § 2) is required and an adjustment of the Birkhoff–Langer procedure ensues, yielding a successful asymptotic treatment.

In § 2, we transform the eigenvalue problem into a first-order ordinary differential equation containing a large parameter.

In § 3, we derive asymptotic estimates of eigenvalues and state the main results.

In § 4, a higher order asymptotic estimation is carried out to help compare the structural and viscous damping effects.

Finally in § 5, we furnish proofs which rigorize the formal asymptotic expansions.

2. Transforming the eigenvalue problem into a first-order system with a large parameter. We consider the eigenvalue problem corresponding to equation (1.11). Let us again assume that the energy-conserving *hinged* boundary conditions (1.2) are in effect. They are imposed here for the sole purpose of simplifying presentations. (Other types of boundary conditions such as the clamped, roller-supported, free, and dissipative conditions can be treated along similar lines of estimation as given here and in [4] and [9], but the calculations usually are much lengthier and more tedious; cf. Remark

2.1.) Using (1.3) and dividing by m, we obtain

$$(2.1)\quad \begin{aligned} &a^4\phi^{(4)}(x)-2\lambda\gamma_1(x)\phi''(x)+2\lambda\gamma_2(x)\phi(x)+\lambda^2\phi(x)=0, \qquad 0\leqq x\leqq \pi,\\ &\phi(0)=\phi''(0)=\phi(\pi)=\phi''(\pi)=0, \end{aligned}$$

where as before $a\equiv(EI/m)^{1/4}$.

Let

$$(2.2)\qquad \Phi(x)=\begin{bmatrix} \phi(x)+\lambda^{-1}a^2\phi''(x)\\ -\phi(x)+\lambda^{-1}a^2\phi''(x)\end{bmatrix}.$$

Then (2.1) is equivalent to

$$(2.3)\quad \begin{aligned} &\Phi''(x)=[\lambda(A+B(x))+C(x)]\Phi(x),\\ &M\Phi(0)=0, \qquad M\Phi(\pi)=0, \end{aligned}$$

where

$$(2.4)\quad \begin{aligned} &A=a^{-2}\begin{bmatrix} 0 & 1\\ -1 & 0\end{bmatrix}, \qquad B(x)=a^{-4}\cdot\gamma_1(x)\begin{bmatrix} 1 & 1\\ 1 & 1\end{bmatrix},\\ &C(x)=a^{-2}\cdot\gamma_2(x)\begin{bmatrix} -1 & 1\\ -1 & 1\end{bmatrix}, \qquad M=\begin{bmatrix} 1 & -1\\ 1 & 1\end{bmatrix}. \end{aligned}$$

To further reduce (2.3) to a first-order system, we let

$$(2.5)\qquad \rho^2=\lambda, \qquad Y(x)=\begin{bmatrix} \Phi(x)\\ \rho^{-1}\Phi'(x)\end{bmatrix},$$

resulting in

$$(2.6)\quad \begin{aligned} &Y'(x)=[\rho(\mathscr{A}+\mathscr{B}(x))+\rho^{-1}\mathscr{C}(x)]Y(x), \qquad 0\leqq x\leqq\pi,\\ &\begin{bmatrix} M & 0\\ 0 & 0\end{bmatrix}Y(0)=0, \qquad \begin{bmatrix} M & 0\\ 0 & 0\end{bmatrix}Y(\pi)=0, \end{aligned}$$

where

$$(2.7)\qquad \mathscr{A}=\begin{bmatrix} 0 & I_2\\ A & 0\end{bmatrix}, \quad \mathscr{B}(x)=\begin{bmatrix} 0 & 0\\ B(x) & 0\end{bmatrix}, \quad \mathscr{C}(x)=\begin{bmatrix} 0 & 0\\ C(x) & 0\end{bmatrix}.$$

Remark 2.1. From the very definitions of $\Phi(x)$ in (2.2) and $Y(x)$ in (2.5), it is easy to see that any set of boundary conditions involving $\phi^{(j)}(0)$, $\phi^{(j)}(\pi)$, $0\leqq j\leqq 3$, two at each end, can be expressed as

$$\mathscr{M}_0Y(0)=0, \qquad \mathscr{M}_\pi Y(\pi)=0,$$

where $\mathscr{M}_0$ and $\mathscr{M}_\pi$ are 4×2 constant matrices. The case of hinged boundary conditions (1.2) results in perhaps the simplest such matrices

$$\mathscr{M}_0=\mathscr{M}_\pi=\begin{bmatrix} M & 0_2\\ 0_2 & 0_2\end{bmatrix} \qquad 0_2 \text{ is the zero } 2\times 2 \text{ matrix.}$$

This is naturally so because the hinged boundary conditions correspond to the "square root" model [5], and it is also observed in (2.2) that only $\phi(x)$ and $\phi''(x)$ (the "square root") appear in $\Phi(x)$. Actually in (2.3) we can just write the boundary conditions as

$$\Phi(0)=0, \qquad \Phi(\pi)=0.$$

To give a presentation in the *general* context, we keep the matrix M here.

We now diagonalize the dominant coefficient matrix $\mathcal{A}+\mathcal{B}(x)$. Let

$$b(x)\equiv i\frac{\sqrt{a^4-\gamma_1^2(x)}}{a^2+\gamma_1(x)},\qquad \xi_1(x)\equiv\frac{[\gamma_1(x)+i\sqrt{a^4-\gamma_1^2(x)}]^{1/2}}{a^2},$$

$$\xi_2(x)\equiv\frac{[\gamma_1(x)-i\sqrt{a^4-\gamma_1^2(x)}]^{1/2}}{a^2},$$

$$\text{(2.8)}\qquad Q(x)=\left[\begin{array}{cc:cc} 1 & 1 & 1 & 1 \\ b(x) & -b(x) & -b(x) & b(x) \\ \hdashline \xi_1(x) & -\xi_2(x) & \xi_2(x) & -\xi_1(x) \\ b(x)\xi_1(x) & b(x)\xi_2(x) & -b(x)\xi_2(x) & -b(x)\xi_1(x) \end{array}\right]$$

$$\equiv\begin{bmatrix} Q_{11}(x) & Q_{12}(x) \\ Q_{21}(x) & Q_{22}(x)\end{bmatrix},$$

where $Q_{ij}(x)$, $1\leqq i,j\leqq 2$, are all 2×2 matrices. Then

$$Q^{-1}(x)=\begin{bmatrix} \frac{1}{4} & \frac{1}{4b(x)} & \frac{1}{4\xi_1(x)} & \frac{1}{4b(x)\xi_1(x)} \\ \frac{1}{4} & -\frac{1}{4b(x)} & -\frac{1}{4\xi_2(x)} & \frac{1}{4b(x)\xi_2(x)} \\ \frac{1}{4} & -\frac{1}{4b(x)} & \frac{1}{4\xi_2(x)} & -\frac{1}{4b(x)\xi_2(x)} \\ \frac{1}{4} & \frac{1}{4b(x)} & -\frac{1}{4\xi_1(x)} & -\frac{1}{4b(x)\xi_1(x)} \end{bmatrix},$$

so

$$\text{(2.9)}\quad Q^{-1}(x)[\mathcal{A}+\mathcal{B}(x)]Q(x)=\begin{bmatrix} \xi_1(x) & & 0 & \\ & -\xi_2(x) & & \\ & & \xi_2(x) & \\ & 0 & & -\xi_1(x) \end{bmatrix}\equiv R(x)$$

is now diagonalized. Let

$$\text{(2.10)}\qquad Y(x)=Q(x)Z(x).$$

Then the differential equation in (2.6) becomes

$$\text{(2.11)}\qquad \begin{aligned} Z'(x)&=\rho Q^{-1}(x)[\mathcal{A}+\mathcal{B}(x)]Q(x)Z(x)-Q^{-1}(x)Q'(x)Z(x)\\ &\quad+\rho^{-1}Q^{-1}(x)\mathcal{C}(x)Q(x)Z(x)\\ &=[\rho R(x)+S(x)+\rho^{-1}T(x)]Z(x), \end{aligned}$$

where

$$\text{(2.12)}\qquad S(x)\equiv -Q^{-1}(x)Q'(x),\qquad T(x)\equiv Q^{-1}(x)\mathcal{C}(x)Q(x),$$

subject to boundary conditions

$$\text{(2.13)}\quad \begin{bmatrix} MQ_{11}(0) & MQ_{12}(0) \\ 0 & 0 \end{bmatrix}Z(0)=0,\qquad \begin{bmatrix} MQ_{11}(\pi) & MQ_{12}(\pi) \\ 0 & 0 \end{bmatrix}Z(\pi)=0.$$

Let us assume that the damping coefficient functions satisfy

$$\text{(2.14)}\qquad \gamma_1,\gamma_2\in C^1([0,\pi]).$$

Consequently from (2.12), we have

$$S \in C^0_{n\times n}([0,\pi]), \qquad T \in C^1_{n\times n}([0,\pi]). \tag{2.15}$$

Following [1]-(3), we now seek a *matrix solution* Z of (2.11) which has a formal asymptotic expansion

$$Z(x) = [P_0(x) + \rho^{-1}P_1(x) + \rho^{-2}P_2(x) + \cdots]\, e^{\rho\Gamma(x)}, \tag{2.16}$$

where

$$\Gamma(x) = \int_0^x R(u)\, du, \tag{2.17}$$

and $P_0(x), P_1(x), \cdots$, are $n\times n$ matrix functions satisfying

$$\begin{aligned}
&\rho^1\colon && P_0R = RP_0\\
&\rho^0\colon && P_0' + P_1R = RP_1 + SP_0\\
&\rho^{-1}\colon && P_1' + P_2R = RP_2 + TP_0 + SP_1\\
& && \vdots\\
&\rho^{-j}\colon && P_j' + P_{j+1}R = RP_{j+1} + TP_{j-1} + SP_j\\
& && \vdots
\end{aligned} \tag{2.18}$$

Remark 2.2. Although (2.11) appears slightly more general than does (19) in [3, p. 71] (due to the presence of the extra term $\rho^{-1}T(x)$ in (2.11)), the asymptotic expansion will go through because the matrix $\mathcal{A} + \mathcal{B}(x)$ has four *distinct* eigenvalues $\pm\xi_1(x)$ and $\pm\xi_2(x)$ on $[0,\pi]$, cf. (2.9). The distinctness of eigenvalues of the matrix $R(x)$ (or equivalently, $\mathcal{A} + \mathcal{B}(x)$) is essential to the applicability of the ansatz (2.16).

Let us denote

$p_{ij}^{(k)}$: the (i,j)-entry of the matrix P_k,

σ_{ij}, τ_{ij}: the (i,j)-entry of matrices S and T, respectively.

$$\zeta_1(x) = \xi_1(x), \quad \zeta_2(x) = -\xi_2(x), \quad \zeta_3(x) = \xi_2(x), \quad \zeta_4(x) = -\xi_1(x). \tag{2.19}$$

Using the arguments as in [3, p. 75], we get

$$\begin{aligned}
&p_{ij}^{(0)}(x) = 0 \quad \text{if } i\neq j, \quad 1\leqq i, \quad j\leqq 4,\\
&p_{jj}^{(0)}(x) = f_j^{(0)} \exp\left[\int_0^x \sigma_{jj}(u)\, du\right], \qquad j=1,2,3,4, \quad f_j \text{ are constants;}
\end{aligned} \tag{2.20}$$

$$\begin{aligned}
&p_{ij}^{(1)}(x) = \frac{\sigma_{ij}(x)p_{jj}^{(0)}(x)}{\zeta_j(x)-\zeta_i(x)} \quad \text{if } i\neq j, \quad 1\leqq i, \quad j\leqq 4,\\
&p_{jj}^{(1)'}(x) - \sum_{k=1}^4 \sigma_{jk}(x)p_{kj}^{(1)}(x) - \sum_{k=1}^4 \tau_{jk}p_{kj}^{(0)}(x) = 0, \qquad k=1,2,3,4
\end{aligned} \tag{2.21}$$

and for $l\geqq 2$,

$$\begin{aligned}
&p_{ij}^{(l)}(x) = \frac{-p_{ij}^{(l-1)'}(x) + \sum_{k=1}^4 [\sigma_{ik}(x)p_{kj}^{(l-1)}(x) + \tau_{ik}(x)p_{kj}^{(l-2)}(x)]}{\zeta_j(x)-\zeta_i(x)},\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad i\neq j, \quad 1\leqq i, \quad j\leqq 4,\\
&p_{jj}^{(l)'}(x) - \sum_{k=1}^4 \sigma_{jk}(x)p_{kj}^{(l)} - \sum_{k=1}^4 \tau_{jk}(x)p_{kj}^{(l-1)}(x) = 0, \qquad i\neq j, \quad 1\leqq i, \quad j\leqq 4.
\end{aligned} \tag{2.22}$$

Note that the second equation in (2.20) and (2.21) can be solved by a quadrature. Thus all the matrices $P_i(x)$, $i=0,1,2,\cdots$, can be formally determined in a recursive way.

3. Asymptotic estimation of eigenfrequencies. We first determine the leading term $P_0(x)$ by (2.20), wherein the diagonal entries $\sigma_{jj}(x)$ of S are easily found to be

$$\sigma_{11}(x)=-\frac{1}{4}\left(\frac{b'(x)}{b(x)}+\frac{\xi_1'(x)}{\xi_1(x)}+\frac{b'(x)\xi_1(x)+b(x)\xi_1'(x)}{b(x)\xi_1(x)}\right)=-\frac{1}{2}\frac{[b(x)\xi_1(x)]'}{b(x)\xi_1(x)},$$

$$\sigma_{22}(x)=-\frac{1}{4}\left(\frac{b'(x)}{b(x)}+\frac{\xi_2'(x)}{\xi_2(x)}+\frac{b'(x)\xi_2(x)+b(x)\xi_2'(x)}{b(x)\xi_2(x)}\right)=-\frac{1}{2}\frac{[b(x)\xi_2(x)]'}{b(x)\xi_2(x)},$$

$$\sigma_{33}(x)=\sigma_{22}(x),\qquad \sigma_{44}(x)=\sigma_{11}(x).$$

Therefore

$$\begin{aligned}
p_{11}^{(0)}(x)&=f_1^{(0)}\cdot\exp\left\{-\frac{1}{2}\int_0^x\frac{[b(u)\xi_1(u)]'}{b(u)\xi_1(u)}\,du\right\}=f_1^{(0)}\cdot\sqrt{\frac{b(0)\xi_1(0)}{b(x)\xi_1(x)}},\\
p_{22}^{(0)}(x)&=f_2^{(0)}\cdot\exp\left\{-\frac{1}{2}\int_0^x\frac{[b(u)\xi_2(u)]'}{b(u)\xi_2(u)}\,du\right\}=f_2^{(0)}\cdot\sqrt{\frac{b(0)\xi_2(0)}{b(x)\xi_2(x)}},\\
p_{33}^{(0)}(x)&=f_3^{(0)}\cdot\sqrt{\frac{b(0)\xi_2(0)}{b(x)\xi_2(x)}},\\
p_{44}^{(0)}(x)&=f_4^{(0)}\sqrt{\frac{b(0)\xi_1(0)}{b(x)\xi_1(x)}}.
\end{aligned}\tag{3.1}$$

The constants $\tilde{f}_j^{(0)}$, $j=1,2,3,4$, can be chosen freely as long as they are nonzero. We simply choose them equal to one so that

(3.2)

$$P_0(x)=\operatorname{diag}\left[\left(\frac{b(0)\xi_1(0)}{b(x)\xi_1(x)}\right)^{1/2},\left(\frac{b(0)\xi_2(0)}{b(x)\xi_2(x)}\right)^{1/2},\left(\frac{b(0)\xi_2(0)}{b(x)\xi_2(x)}\right)^{1/2},\left(\frac{b(0)\xi_1(0)^{1/2}}{b(x)\xi_1(x)}\right)^{1/2}\right].$$

We now formally write (2.16) as

$$\begin{aligned}
Z(x)&=\left[P_0(x)+\frac{P_1(x)}{\rho}+\frac{P_2(x)}{\rho^2}+\cdots\right]e^{\rho\Gamma(x)}\\
&\equiv\left[P_0(x)+\frac{\Psi(x,\rho)}{\rho}\right]e^{\rho\Gamma(x)},
\end{aligned}\tag{3.3}$$

where

$$\begin{aligned}
\Psi(x,\rho)&=P_1(x)+\rho^{-1}P_2(x)+\cdots\\
&\equiv[\psi_{ij}(x,\rho)]_{4\times4},\qquad 1\leqq i,\quad j\leqq 4.
\end{aligned}\tag{3.4}$$

Since $Y(x)$ as defined in (2.4), (2.5) is a *vector solution*, and since $Z(x)$ is the *fundamental matrix solution* of (2.11), the transformation (2.10) will actually give us

$$Y(x)=Q(x)Z(x)\vec{c},\tag{3.5}$$

where $\vec{c}$ is a 4×1 constant vector with entries c_1, c_2, c_3, and c_4 in $\mathbb{C}$. Thus (2.13) becomes

$$\begin{bmatrix}MQ_{11}(0)&MQ_{12}(0)\\0&0\end{bmatrix}Z(0)\vec{c}=0,\qquad\begin{bmatrix}MQ_{11}(\pi)&MQ_{12}(\pi)\\0&0\end{bmatrix}Z(\pi)\vec{c}=0.\tag{3.6}$$

From (2.8), (3.2), (3.3), we obtain from the boundary conditions in (2.13) that

$$
(3.7)\quad
\begin{bmatrix} M & 0_2 \\ 0_2 & M \end{bmatrix}
\cdot
\begin{bmatrix}
1+\frac{\tilde{\psi}_{11}(\rho)}{\rho} & 1+\frac{\tilde{\psi}_{12}(\rho)}{\rho} & 1+\frac{\tilde{\psi}_{13}(\rho)}{\rho} & 1+\frac{\tilde{\psi}_{14}(\rho)}{\rho} \\
b(0)+\frac{\tilde{\psi}_{21}(\rho)}{\rho} & -b(0)+\frac{\tilde{\psi}_{22}(\rho)}{\rho} & -b(0)+\frac{\tilde{\psi}_{23}(\rho)}{\rho} & b(0)+\frac{\tilde{\psi}_{24}(\rho)}{\rho} \\
K_1E_1(\rho) & K_2E_2(\rho) & K_2E_3(\rho) & K_1E_4(\rho) \\
\quad +\frac{E_1(\rho)\tilde{\psi}_{31}(\rho)}{\rho} & \quad +\frac{E_2(\rho)\tilde{\psi}_{32}(\rho)}{\rho} & \quad +\frac{E_3(\rho)\tilde{\psi}_{33}(\rho)}{\rho} & \quad +\frac{E_4(\rho)\tilde{\psi}_{34}(\rho)}{\rho} \\
b(\pi)K_1E_1(\rho) & -b(\pi)K_2E_2(\rho) & -b(\pi)K_2E_3(\rho) & b(\pi)K_1E_4(\rho) \\
\quad +\frac{E_1(\rho)\tilde{\psi}_{41}(\rho)}{\rho} & \quad +\frac{E_2(\rho)\tilde{\psi}_{42}(\rho)}{\rho} & \quad +\frac{E_3(\rho)\tilde{\psi}_{43}(\rho)}{\rho} & \quad +\frac{E_4(\rho)\tilde{\psi}_{44}(\rho)}{\rho}
\end{bmatrix}
\cdot \vec{c}
\equiv \mathbb{M}\Lambda(\rho)\vec{c} = 0,
$$

where

$$
(3.8)\quad
\begin{aligned}
&K_1 \equiv [b(0)\xi_1(0)/b(0)\xi_1(\pi)]^{1/2}, \qquad K_2 = [b(0)\xi_2(0)/b(\pi)\xi_2(\pi)]^{1/2}, \\
&\Gamma_1(\pi) = \int_0^\pi \xi_1(x)\,dx, \qquad \Gamma_2(\pi) = \int_0^\pi \xi_2(x)\,dx, \\
&E_1(\rho) = e^{\rho\Gamma_1(\pi)}, \quad E_2(\rho) = e^{-\rho\Gamma_2(\pi)}, \quad E_3(\rho) = e^{\rho\Gamma_2(\pi)}, \quad E_4(\rho) = e^{-\rho\Gamma_1(\pi)}, \\
&\tilde{\psi}_{1k}(\rho) = \sum_{i=1}^4 \psi_{ik}(0,\rho), \quad \tilde{\psi}_{2k}(\rho) = b(0)\sum_{i=1}^4 (-1)^{[i/2]}\psi_{ik}(0,\rho), \\
&\tilde{\psi}_{3k}(\rho) = \sum_{i=1}^4 \psi_{ik}(\pi,\rho), \quad \tilde{\psi}_{4k}(\rho) = b(\pi)\sum_{i=1}^4 (-1)^{[i/2]}\psi_{ik}(\pi,\rho), \qquad k = 1,2,3,4.
\end{aligned}
$$

It is easy to see that K_1, K_2, $\Gamma_1(\pi)$, and $\Gamma_2(\pi)$ are bounded. From [3], we know that $\tilde{\psi}_{jk}(\rho)$, $1 \leqq j, k \leqq 4$, are uniformly bounded for $|\rho|$ large. ρ is an eigenvalue of (2.11) and (2.13) if and only if the determinant of the product matrix in (3.7) is zero. But the matrix $\mathbb{M}$ in (3.7) is invertible, and the constant vector $\vec{c}$ is nontrivial, therefore

$$
(3.9)\qquad \det \Lambda(\rho) = 0.
$$

First let us determine the roots of the simpler equation

$$
(3.10)\quad \det \begin{bmatrix}
1 & 1 & 1 & 1 \\
b(0) & -b(0) & -b(0) & b(0) \\
K_1E_1 & K_2E_2 & K_2E_3 & K_1E_4 \\
b(\pi)K_1E_1 & -b(\pi)K_2E_2 & -b(\pi)K_2E_3 & b(\pi)K_1E_4
\end{bmatrix} \equiv \det \Lambda_0 = 0,
$$

obtained by neglecting $\mathcal{O}(\rho^{-1})$ terms in $\Lambda(\rho)$. This determinant is easy to expand, giving

$$
(3.11)\qquad \det \Lambda_0 = 4b(0)b(\pi)K_1K_2(E_1 - E_4)(E_3 - E_2) = 0.
$$

By (1.12), $b(0)$, $b(\pi)$, K_1, and K_2 are all nonzero. Therefore

$$
(3.12)\qquad E_1 - E_4 = e^{\rho\Gamma_1(\pi)} - e^{-\rho\Gamma_1(\pi)} = 0,
$$

or

$$E_3 - E_2 = e^{\rho\Gamma_2(\pi)} - e^{-\rho\Gamma_2(\pi)} = 0. \tag{3.13}$$

The roots ρ of (3.12) are determined by

$$e^{2\rho\Gamma_1(\pi)} = 1,$$

$$\begin{aligned}\rho &= \frac{1}{2\Gamma_1(\pi)} \cdot i2k\pi, \qquad k \in \mathbb{Z} \\ &= i \cdot \frac{a^2 k\pi}{\int_0^\pi [\gamma_1(x) + i\sqrt{a^4 - \gamma_1^2(x)}]\, dx}.\end{aligned} \tag{3.14}$$

Noting that

$$[\gamma_1(x) + i\sqrt{a^4 - \gamma_1^2(x)}]^{1/2} = \frac{1}{\sqrt{2}}[\sqrt{a^2 + \gamma_1(x)} + i\sqrt{a^2 - \gamma_1(x)}],$$

we obtain from (3.14) that

$$\begin{aligned}\rho_1(k) &\equiv i \frac{\sqrt{2}\, ka^2\pi}{\int_0^\pi \sqrt{a^2 + \gamma_1(x)}\, dx + i\int_0^\pi \sqrt{a^2 - \gamma_1(x)}\, dx} \\ &= \frac{\beta + i\alpha}{\alpha^2 + \beta^2}(\sqrt{2}\, a^2\pi) \cdot k, \qquad k \in \mathbb{Z},\end{aligned} \tag{3.15}$$

where

$$\alpha \equiv \int_0^\pi \sqrt{a^2 + \gamma_1(x)}\, dx, \qquad \beta \equiv \int_0^\pi \sqrt{a^2 - \gamma_1(x)}\, dx. \tag{3.16}$$

Similarly, from (3.13) we obtain

$$\rho_2(k) = \frac{-\beta + i\alpha}{\alpha^2 + \beta^2}(\sqrt{2}\, a^2\pi) \cdot k, \qquad k \in \mathbb{Z}. \tag{3.17}$$

Let

$$G_{ik} = \{\rho \in \mathbb{C} \mid {}^{-1 \leqq \operatorname{Re}(\rho - \rho_i(k)) \leqq 1}_{-1 \leqq \operatorname{Im}(\rho - \rho_i(k)) \leqq 1}\}, \qquad i = 1, 2, \quad |k| \geqq k_0,$$

$$G_i^+ = \bigcup_{k=k_0}^{+\infty} G_{ik}, \quad G_i^- = \bigcup_{k=-k_0}^{-\infty} G_{ik}, \quad i = 1, 2,$$

$$G_i = G_i^+ \cup G_i^-,$$

where k_0 is a large positive integer. Then we have the following lemma.

LEMMA 3.1.

$$\det \Lambda(\rho) = \det \Lambda_0 + \rho^{-1}(g_1(\rho)E_3(\rho) + g_2(\rho)), \qquad \rho \in G_1^+, \tag{3.18}$$

$$\det \Lambda(\rho) = \det \Lambda_0 + \rho^{-1}(g_3(\rho)E_2(\rho) + g_4(\rho)), \qquad \rho \in G_1^-, \tag{3.18$'$}$$

and

$$\det \Lambda(\rho) = \det \Lambda_0 + \rho^{-1}(g_5(\rho)E_1(\rho) + g_6(\rho)), \qquad \rho \in G_2^+, \tag{3.19}$$

$$\det \Lambda(\rho) = \det \Lambda_0 + \rho^{-1}(g_7(\rho)E_4(\rho) + g_8(\rho)), \qquad \rho \in G_2^-, \tag{3.19$'$}$$

where $g_i(\rho)$, $i = 1, 2, \cdots, 8$ *is uniformly bounded.*

The proofs of Lemmas 3.1 and 3.2 will be given in § 5.

As in [8], we compare the roots of (3.9) with those of (3.10). The results show that the two sequences of roots $\rho_1(k)$, $\rho_2(k)$ constitute a first-order asymptotic approximation of roots of (3.9) in the following sense.

LEMMA 3.2. *Let $\rho_1(k)$, $\rho_2(k)$, $k \in \mathbb{Z}$, be the roots of* (3.10) *and let $\tilde{\rho}$ be a root of* (3.9). *Then there exists a bound $B > 0$ such that for any $\tilde{\rho}$: $|\tilde{\rho}| \geqq B$, there exists some $\tilde{k} \in \mathbb{Z}$ such that*

$$|\tilde{\rho} - \rho_j(\tilde{k})| \leqq 1, \qquad j = 1 \text{ or } 2, \tag{3.20}$$

where $\tilde{k}$ is dependent only on $\tilde{\rho}$.

Consequently, any solution $\tilde{\rho}$ of (3.9) satisfies

$$\tilde{\rho} = \rho_j(k) + \mathcal{O}(1), \qquad j = 1, 2, \quad |k| \text{ large.} \tag{3.21}$$

From (3.21), we can now determine the asymptotic distribution pattern of the eigenfrequencies λ of (1.11) and (1.2). From (2.5), (3.15), (3.17),

$$\begin{aligned}\lambda = \rho^2 &= [\rho_j(k) + \mathcal{O}(1)]^2 \qquad k = 1 \text{ or } 2, |k| \text{ large},\\ &= \frac{(-\alpha^2 + \beta^2) \pm 2i\alpha\beta}{(\alpha^2 + \beta^2)^2}(2a^4\pi^2k^2) + \mathcal{O}(k), \quad \text{“+” for } j = 1, \text{ “−” for } j = 2.\end{aligned}$$

Therefore we see that asymptotically eigenfrequencies λ are located at the vicinity of two rays on the left half plane whose angles with respect to the imaginary axis are

$$\begin{aligned}\theta_j &\equiv \lim_{k\to\infty} \tan^{-1} \frac{(\alpha^2 + \beta^2)^{-2}[(-\alpha^2 + \beta^2)(2a^4\pi^2k^2) + \mathcal{O}(k)]}{(\alpha^2 + \beta^2)^{-2}[(\pm 2\alpha\beta)(2a^4\pi^2k^2) + \mathcal{O}(k)]},\\ &\qquad\qquad \text{“+” for } j = 1, \text{ “−” for } j = 2,\\ &= \pm\tan^{-1}\frac{\beta^2 - \alpha^2}{2\alpha\beta}.\end{aligned} \tag{3.22}$$

We conclude

THEOREM 3.3. *Let $\lambda = \mu + i\nu$, $\mu, \nu \in \mathbb{R}$, be a large eigenfrequency of* (1.11) *and* (1.2). *Then*

$$\lim_{|\lambda|\to\infty} \frac{\mu}{\nu} = \pm\frac{1}{2}\frac{[\int_0^\pi \sqrt{a^2 + \gamma_1(x)}\,dx]^2 - [\int_0^\pi \sqrt{a^2 - \gamma_1(x)}\,dx]^2}{[\int_0^\pi \sqrt{a^2 + \gamma_1(x)}\,dx] \cdot [\int_0^\pi \sqrt{a^2 - \gamma_1(x)}\,dx]}. \tag{3.23}$$

Hence the damping rate μ is asymptotically proportional to the frequency ν, with the constants of proportionality (3.23) *depending only on the function $\gamma_1(x)$ satisfying* (1.12), (2.14).

Remark 3.4. (i) Theorem 3.3 points out that the answer to (Q1) in § 1 is *negative.* The “homogenization” formula (3.23) is much more complex than the one in (Q1).

(ii) Note that when

$$\gamma_1(x) \equiv \gamma_1 \quad \text{a positive constant,} \tag{3.24}$$

then the slopes in (3.23) become

$$\pm\frac{1}{2}\frac{[\pi\sqrt{a^2 + \gamma_1}]^2 - [\pi\sqrt{a^2 - \gamma_1}]^2}{[\pi\sqrt{a^2 + \gamma_1}][\pi\sqrt{a^2 - \gamma_1}]} = \pm\frac{\gamma_1}{\sqrt{a^4 - \gamma_1^2}},$$

consistent with (1.6).

Remark 3.5. If instead of (1.11), we treat (1.11)′, then the eigenvalue problem (2.1) becomes

(2.1)′

$$a^4\phi^{(4)}(x) - 2\lambda\gamma_1(x)\phi''(x) - 2\lambda\gamma_1'(x)\phi'(x) + 2\lambda\gamma_2(x)\phi(x) + \lambda^2\phi(x) = 0, \qquad 0 \leqq x \leqq \pi,$$

$$\phi(0) = \phi''(0) = \phi(\pi) = \phi''(\pi) = 0.$$

Following the same transformation and diagonalization procedures as in (2.1)-(2.11), we will obtain, in lieu of (2.11), an equivalent first-order system

$$Z'(x)=[\rho R(x)+S(x)+\rho^{-1}T(x)]Z(x), \tag{2.11$'$}$$

where

$$S(x)=Q^{-1}(x)\cdot\mathscr{D}(x)Q(x)-Q^{-1}(x)Q'(x),$$

$$\mathscr{D}(x)\equiv\begin{bmatrix}0 & 0\\ 0 & D(x)\end{bmatrix},\qquad D(x)\equiv a^{-2}\gamma_1'(x)\begin{bmatrix}1 & -1\\ 1 & -1\end{bmatrix},$$

and everything else remains unchanged. Since (3.23) in Theorem 3.3 is independent of $S(x)$ or $S(x)$ in (2.11) or (2.11)$'$, we see that the very same structural damping property holds good for (1.11)$'$.

4. A high-order approximation under the assumption of uniform structural damping in order to examine the effects of viscous damping. So far in our analysis of damping, we see that in (1.11) the effect of the term $-2m\gamma_1(x)\partial^3y(x,t)/\partial t\partial x^2$ completely overwhelms that of the viscous $2m\gamma_2(x)\partial y(x,t)/\partial t$. Nevertheless, we are still curious to determine what is the order of magnitude of the (asymptotic) effect of the viscous damping term.

We make a quick observation at (2.11): in

$$Z'(x)=[\rho R(x)+S(x)+\rho^{-1}T(x)]Z(x),$$

only the matrix $\rho^{-1}T(x)$ depends on the viscous coefficient $\gamma_2(x)$, with magnitude $\mathscr{O}(\rho^{-1})$, whereas the matrix $S(x)$, depending on $\gamma_1(x)$ and $\gamma_1'(x)$, still has a larger order of magnitude $\mathscr{O}(1)$. Therefore even if we go to the second-order approximation of $Z(x)$, i.e., to obtain $P_1(x)$ in (2.16), the asymptotic effect contributed by viscosity $\gamma_2(x)$ still would be overshadowed by $S(x)$. In order to satisfy our probing curiosity and to avoid lengthy calculations, we assume

$$S(x)\equiv 0,$$

resulting in a simpler equation

$$Z'(x)=[\rho R(x)+\rho^{-1}T(x)]Z(x). \tag{4.1}$$

This is equivalent to assuming (3.24) for model (1.11) or (1.11)$'$. (The reader can also see from (2.22) that if $S(x)\not\equiv 0$, the solution of $P_1(x)$ is a great deal more complicated in general because it may not be in diagonal form and lack closed form solutions, causing a severe obstacle in obtaining higher order approximations.)

Under the uniform structural damping assumption (3.24), now we can obtain $P_1(x)$ explicitly as a diagonal matrix:

$$\begin{aligned}P_1(x)&=\operatorname{diag}\left[\int_0^x\tau_{11}(u)\,du,\int_0^x\tau_{22}(u)\,du,\int_0^x\tau_{33}(u)\,du,\int_0^x\tau_{44}(u)\,du\right]\\&=\frac{b^2-1}{4a^2b}\operatorname{diag}\left[\frac{1}{\xi_1}\eta(x),-\frac{1}{\xi_2}\eta(x),\frac{1}{\xi_2}\eta(x),-\frac{1}{\xi_1}\eta(x)\right],\end{aligned}$$

where

$$b=i\frac{\sqrt{a^4-\gamma_1^2}}{a^2+\gamma_1}\quad\text{(from (2.8))},$$

ξ_1, ξ_2 as in (2.8) formerly, now are constant through (3.24),

$$\eta(x)\equiv\int_0^x\gamma_2(u)\,du.$$

The very fact that $P_1(x)$ is a diagonal matrix crucially facilitates the tractability of our problem. Also, from (3.2) and (2.9), (2.17),

$$(4.2)\qquad P_0(x)=I_4,\qquad \Gamma(x)=\operatorname{diag}[\xi_1x,-\xi_2x,\xi_2x,-\xi_1x].$$

From the ansatz (2.16), using (4.2), we can rewrite

$$Z(x)=[I+\rho^{-2}\hat{\Psi}(x,\rho)]E(x,\rho),$$

where

$$\begin{aligned}E(x,\rho)&\equiv[I+\rho^{-1}P_1(x)]\,e^{\rho\Gamma(x)}\\ &=\operatorname{diag}\left(\left(1+\frac{b^2-1}{4a^2b\xi_1\rho}\,\eta(x)\right)\exp(\rho\xi_1x),\left(1+\frac{b^2-1}{4a^2b\xi_2\rho}\,\eta(x)\right)\exp(-\rho\xi_2x),\right.\\ &\qquad\left.\cdot\left(1-\frac{b^2-1}{4a^2b\xi_2\rho}\,\eta(x)\right)\exp(\rho\xi_2x),\left(1-\frac{b^2-1}{4a^2b\xi_1\rho}\,\eta(x)\right)\exp(-\rho\xi_1x)\right),\end{aligned}\tag{4.3}$$

$$\hat{\Psi}(x,\rho)\equiv[P_2(x)+\rho^{-1}P_3(x)+\cdots+\rho^{-n}P_{n+2}(x)+\cdots][I+\rho^{-1}P_1(x)]^{-1}.$$

The boundedness and holomorphic properties of this function $\hat{\Psi}(x,\rho)$ are similar to those of $\Psi(x,\rho)$ in (3.4) so we will not repeat the discussions here. The boundary conditions in (3.6) require that ρ satisfies

$$(4.4)\qquad \det\{\tilde{\Lambda}(\rho)+\rho^{-2}\tilde{\hat{\Psi}}(\rho)\}=0,$$

where

$$\tilde{\Lambda}(\rho)=\begin{bmatrix}1&1&1&1\\ b&-b&-b&b\\ \left[1+\frac{b^2-1}{4a^2b\xi_1\rho}\eta(\pi)\right]&\left[1+\frac{b^2-1}{4a^2b\xi_2\rho}\eta(\pi)\right]&\left[1-\frac{b^2-1}{4a^2b\xi_2\rho}\eta(\pi)\right]&\left[1-\frac{b^2-1}{4a^2b\xi_1\rho}\eta(\pi)\right]\\ \cdot\exp(\rho\xi_1\pi)&\cdot\exp(-\rho\xi_2\pi)&\cdot\exp(\rho\xi_2\pi)&\cdot\exp(-\rho\xi_1\pi)\\ b\left[1+\frac{b^2-1}{4a^2b\xi_1\rho}\eta(\pi)\right]&-b\left[1+\frac{b^2-1}{4a^2b\xi_2\rho}\eta(\pi)\right]&-b\left[1-\frac{b^2-1}{4a^2b\xi_2\rho}\eta(\pi)\right]&b\left[1-\frac{b^2-1}{4a^2b\xi_1\rho}\eta(\pi)\right]\\ \cdot\exp(\rho\xi_1\pi)&\cdot\exp(-\rho\xi_2\pi)&\cdot\exp(\rho\xi_2\pi)&\cdot\exp(-\rho\xi_1\pi)\end{bmatrix}$$

and

$$\tilde{\hat{\Psi}}(\rho)=\begin{bmatrix}\tilde{\hat{\psi}}_{11}(\rho)&\tilde{\hat{\psi}}_{12}(\rho)&\tilde{\hat{\psi}}_{13}(\rho)&\tilde{\hat{\psi}}_{14}(\rho)\\ \tilde{\hat{\psi}}_{21}(\rho)&\tilde{\hat{\psi}}_{22}(\rho)&\tilde{\hat{\psi}}_{23}(\rho)&\tilde{\hat{\psi}}_{24}(\rho)\\ \tilde{\hat{\psi}}_{31}(\rho)\exp(\rho\xi_1\pi)&\tilde{\hat{\psi}}_{32}(\rho)\exp(-\rho\xi_2\pi)&\tilde{\hat{\psi}}_{33}(\rho)\exp(\rho\xi_2\pi)&\tilde{\hat{\psi}}_{34}(\rho)\exp(-\rho\xi_1\pi)\\ \tilde{\hat{\psi}}_{41}(\rho)\exp(\rho\xi_1\pi)&\tilde{\hat{\psi}}_{42}(\rho)\exp(-\rho\xi_2\pi)&\tilde{\hat{\psi}}_{43}(\rho)\exp(\rho\xi_2\pi)&\tilde{\hat{\psi}}_{44}(\rho)\exp(-\rho\xi_1\pi)\end{bmatrix}\tag{4.5}$$

with

$$\tilde{\hat{\psi}}_{1k}(\rho)=\sum_{i=1}^{4}\hat{\psi}_{ik}(0,\rho),$$

$$\tilde{\hat{\psi}}(\rho)=b\sum_{i=1}^{4}(-1)^{[i/2]}\hat{\psi}_{ik}(0,\rho),$$

$$\tilde{\hat{\psi}}_{3k}(\rho)=\sum_{i=1}^{4}\hat{\psi}_{ik}(\pi,\rho),$$

$$\tilde{\hat{\psi}}_{4k}(\rho)=b\sum_{i=1}^{4}(-1)^{[i/2]}\hat{\psi}_{ik}(\pi,\rho)$$

for $k=1,2,3,4$.

The approximation of ρ is done by dropping the $\mathcal{O}(\rho^{-2})$ terms in (4.4):

$$\begin{aligned} 0 &= \det \tilde{\Lambda}(\rho) \\ &= 4b^2\left[\left(1+\frac{b^2-1}{4a^2b\xi_1\rho}\,\eta(\pi)\right)e^{\rho\xi_1\pi}-\left(1-\frac{b^2-1}{4a^2b\xi_1\rho}\,\eta(\pi)\right)e^{-\rho\xi_1\pi}\right] \\ &\quad\cdot\left[\left(1-\frac{b^2-1}{4a^2b\xi_2\rho}\,\eta(\pi)\right)e^{\rho\xi_2\pi}-\left(1+\frac{b^2-1}{4a^2b\xi_2\rho}\,\eta(\pi)\right)e^{-\rho\xi_2\pi}\right]. \end{aligned} \tag{4.6}$$

This is of a higher order approximation to ρ than that given in § 3. Therefore ρ satisfies either

$$\left[1+\frac{b^2-1}{4a^2b\xi_1\rho}\,\eta(\pi)\right]e^{\rho\xi_1\pi}-\left[1-\frac{b^2-1}{4a^2b\xi_1\rho}\,\eta(\pi)\right]e^{-\rho\xi_1\pi}=0 \tag{4.7}$$

or

$$\left[1-\frac{b^2-1}{4a^2b\xi_2\rho}\,\eta(\pi)\right]e^{\rho\xi_2\pi}-\left[1+\frac{b^2-1}{4a^2b\xi_2\rho}\right]e^{\rho\xi_2\pi}=0. \tag{4.8}$$

Note that (4.7) is obtainable from (4.8) by the formal change of variable

$$\xi_2=-\xi_1, \tag{4.9}$$

therefore we need only solve (4.7), where from

$$\begin{aligned} e^{2\pi\xi_1\rho} &= \left(1-\frac{(b^2-1)\eta(\pi)}{4a^2b\xi_1\rho}\right)\Big/\left(1+\frac{(b^2-1)\eta(\pi)}{4a^2b\xi_1\rho}\right)\equiv\omega(\rho), \\ \rho &= \frac{1}{2\pi\xi_1}\ln\omega(\rho)+i\frac{k}{\xi_1}, \qquad k\in\mathbb{Z}. \end{aligned} \tag{4.10}$$

Since

(4.11)

$$\ln\omega(\rho)=-\frac{(b^2-1)\eta(\pi)}{2a^2b\xi_1}\frac{1}{\rho}+2\cdot\left[\frac{(b^2-1)\eta(\pi)}{4a^2b\xi_1}\right]^2\frac{1}{\rho^2}+\cdots=\mathcal{O}(\rho^{-1}), \qquad |\rho| \text{ large},$$

so

$$\frac{k}{\rho}=-i\xi_1+\mathcal{O}(\rho^{-2}), \tag{4.12}$$

we have

$$\begin{aligned} \rho^2 &= -\frac{k^2}{\xi_1^2}+\frac{k}{\pi\xi_1^2}\ln\omega(\rho)+\frac{1}{4\pi^2\xi_1^2}\ln^2(\omega(\rho)) \\ &\equiv \mathcal{F}_1+\mathcal{F}_2+\mathcal{F}_3, \end{aligned} \tag{4.13}$$

where

$$\mathcal{F}_1=-\frac{k^2}{\xi_1^2}=-\frac{a^4k^2}{\gamma_1+i\sqrt{a^4-\gamma_1^2}}=-k^2\gamma_1+ik^2\sqrt{a^4-\gamma_1^2},$$

$$
\begin{aligned}
\mathscr{F}_2 &= i\frac{k}{\pi\xi_1^2}\ln \omega(\rho)\\
&= i\frac{k}{\pi\xi_1^2}\left[-\frac{(b^2-1)\eta(\pi)}{2a^2b\xi_1}\frac{1}{\rho}+\mathscr{O}(\rho^{-2})\right] \quad \text{(by (4.11))}\\
&= -i\frac{(b^2-1)\eta(\pi)}{2a^2b\pi\xi_1^3}\frac{k}{\rho}+\mathscr{O}(\rho^{-1})\\
&= -\frac{(b^2-1)\eta(\pi)}{2a^2b\pi\xi_1^2}+\mathscr{O}(\rho^{-1}) \quad \text{(by (4.12))}\\
&= -\frac{\left[-\dfrac{(a^4-\gamma_1^2)}{(a^2+\gamma_1)^2}-1\right]\eta(\pi)}{2a^2\pi\cdot i\cdot\dfrac{(a^4-\gamma_1^2)^{1/2}}{a^2+\gamma_1}\cdot\left[\dfrac{\gamma_1+i(a^4-\gamma_1^2)^{1/2}}{a^4}\right]}+\mathscr{O}(\rho^{-2})\\
&= -\frac{\eta(\pi)}{\pi}-i\frac{\gamma_1}{\sqrt{a^4-\gamma_1^2}}\frac{\eta(\pi)}{\pi}+\mathscr{O}(\rho^{-2})\\
\mathscr{F}_3 &= \mathscr{O}(\rho^{-2}).
\end{aligned}
$$

Combining the above in (4.13), we therefore have

$$
\rho^2=-\left(k^2\gamma_1+\frac{\eta(\pi)}{\pi}\right)+i\left(k^2\sqrt{a^4-\gamma_1^2}-\frac{\gamma_1}{\sqrt{a^4-\gamma_1^2}}\frac{\eta(\pi)}{\pi}\right)+\mathscr{O}\left(\frac{1}{k}\right), \tag{4.14}
$$

$k\in\mathbb{Z}$ is large.

Similarly, (4.8) will lead to

$$
\rho^2=-\left(k^2\gamma_1+\frac{\eta(\pi)}{\pi}\right)-i\left(k^2\sqrt{a^4-\gamma_1^2}-\frac{\gamma_1}{\sqrt{a^4-\gamma_1^2}}\frac{\eta(\pi)}{\pi}\right)+\mathscr{O}\left(\frac{1}{k}\right). \tag{4.15}
$$

Now we return to solve

$$
\det[\tilde{\Lambda}(\sqrt{\lambda})+\lambda^{-1}\tilde{\Psi}(\sqrt{\lambda})]=0, \tag{4.16}
$$

namely, (4.4). By a careful perturbation argument, we can again show that the eigenfrequencies λ of (1.11), (1.2) have the asymptotic expansion

$$
\lambda=-\left(k^2\gamma_1+\frac{\eta(\pi)}{\pi}\right)\pm i\left(k^2\sqrt{a^4-\gamma_1^2}-\frac{\gamma_1}{\sqrt{a^4-\gamma_1^2}}\frac{\eta(\pi)}{\pi}\right)+\mathscr{O}\left(\frac{1}{k}\right), \quad |\lambda| \text{ large.} \tag{4.17}
$$

Therefore we conclude

THEOREM 3.6. *Consider* (1.11) *and* (1.2). *Assume that* $\gamma_1(x)\equiv\gamma_1$ *is constant satisfying* (1.7). *Let* λ *be an eigenfrequency of the system*; *then for* $|\lambda|$ *large,*

$$
\lambda=-\left(k^2\gamma_1+\frac{\eta(\pi)}{\pi}\right)\pm i\left(k^2\sqrt{a^4-\gamma_1^2}-\frac{\gamma_1}{\sqrt{a^4-\gamma_1^2}}\frac{\eta(\pi)}{\pi}\right)+\mathscr{O}\left(\frac{1}{k}\right), \tag{4.18}
$$

$k\in\mathbb{Z}$ *is large.*

Therefore we see that, asymptotically, the contribution of the viscous damping term $2m\gamma_2(x)\partial y/\partial t$ to the total rate of decay

$$
-\left[k^2\gamma_1+\frac{\eta(\pi)}{\pi}\right] \tag{4.19}
$$

is rather insignificant because the contribution $\eta(\pi)/\pi=\int_0^\pi \gamma_2(x)\,dx/\pi$ is fixed, whereas $k^2\gamma_1$ can grow large very fast. If $\gamma_1(x)$ is not constant, then the asymptotic decay rate (4.19) would contain a $\mathcal{O}(k)$ term, which is still of larger order of magnitude than $\eta(\pi)/\pi$. Also, it contributes a constant phase shift

$$-\frac{\gamma_1}{\sqrt{a^4-\gamma_1^2}}\,\frac{\eta(\pi)}{\pi}$$

to the overall (temporal) phase

$$k^2\sqrt{a^4-\gamma_1^2}-\frac{\gamma_1}{\sqrt{a^4-\gamma_1^2}}\,\frac{\eta(\pi)}{\pi}.$$

A special case of (4.18) is when

$$\gamma_1(x)\equiv\gamma_1=0.$$

Then we have

$$\begin{aligned}\lambda&=-\frac{\eta(\pi)}{\pi}\pm i\cdot k^2a^2+\mathcal{O}\left(\frac{1}{k}\right)\\ &=-\frac{1}{\pi}\int_0^\pi \gamma_2(x)\,dx\pm i\cdot k^2a^2+\mathcal{O}\left(\frac{1}{k}\right).\end{aligned}\tag{4.20}$$

This says that the rate of decay is the average of the viscous damping coefficient function $\gamma_2(x)$ on $[0,\pi]$. This is consistent with (1.14) and has also answered (Q2) *affirmatively* which was posed earlier in § 1.

5. Proofs of two technical lemmas.

Proof of Lemma 3.1. We pointed out earlier in § 3 that in $\Lambda(\rho)$ of (3.7), K_1, K_2, $\Gamma_1(\pi)$, and $\Gamma_2(\pi)$ are finite, $\psi_{jk}(\rho)$, $1\leqq j,\ k\leqq 4$, are uniformly bounded for $|\rho|$ large. Let k_0 be a sufficiently large positive integer, and $k\in\mathbb{Z}$, $|k|\geqq k_0$. Let $\rho\in G_{1k}$. Then

$$\begin{aligned}\rho&=\frac{\beta}{\alpha^2+\beta^2}(\sqrt{2}\,a^2\pi)k+\varepsilon_1+i\left[\frac{\alpha}{\alpha^2+\beta^2}(\sqrt{2}\,a^2\pi)k+\varepsilon_2\right]\\ &=\rho_1(k)+(\varepsilon_i+i\varepsilon_2)\qquad\text{(cf. (3.16))},\end{aligned}$$

where $|\varepsilon_1|\leqq 1$, $|\varepsilon_2|\leqq 1$. We have

$$\begin{aligned}|E_1(\rho)|&=|e^{[\rho_1(k)+\varepsilon_1+i\varepsilon_2]\Gamma_1(\pi)}|=|e^{\rho_1(k)\Gamma_1(\pi)}\cdot e^{(\varepsilon_1+i\varepsilon_2)\Gamma-1(\pi)}|\\ &=|e^{(\varepsilon_1+i\varepsilon_2)\Gamma_1(\pi)}|,\qquad(\because\rho_1(k)\Gamma_1(\pi)=ik\pi),\\ |E_4(\rho)|&=|e^{-[\rho_1(k)+\varepsilon_1+i\varepsilon_2]\Gamma_1(\pi)}|=|e^{-(\varepsilon_1+i\varepsilon_2)\Gamma_1(\pi)}|,\\ |E_2(\rho)|&=|e^{-[\rho_1(k)+\varepsilon_1+i\varepsilon_2]\Gamma_2(\pi)}|=|e^{-(\varepsilon_1+i\varepsilon_2)\Gamma_2(\pi)}|\left|\exp\left(-\frac{\beta+i\alpha}{\alpha^2+\beta^2}(\sqrt{2}\,a^2k\pi)(\alpha-i\beta)\right)\right|\\ &=|e^{-(\varepsilon_1+i\varepsilon_2)\Gamma_2(\pi)}|\cdot\left|\exp\left(-\left(\frac{2\alpha\beta}{\alpha^2+\beta^2}+i\frac{\alpha^2-\beta^2}{\alpha^2+\beta^2}\right)\cdot\sqrt{2}\,a^2k\pi\right)\right|,\end{aligned}$$

so

$$|E_2(\rho)|\to+\infty\quad\text{as }k\to+\infty,$$

$$|E_3(\rho)|\to+0\quad\text{as }k\to-\infty.$$

Similarly, we can show that

$$|E_3(\rho)| \to +\infty \quad \text{as } k \to +\infty,$$
$$|E_3(\rho)| \to 0 \quad \text{as } k \to -\infty.$$

Let us look at the case $\rho \in G_{1k}^+$, $k \geqq k_0$. From the above, we see that $E_1(\rho)$, $E_2(\rho)$, and $E_4(\rho)$ are bounded uniformly in k in ρ, while $|E_3(\rho)| \to +\infty$ as $k \to +\infty$. Applying the fact that

$$\det \begin{bmatrix} a_{11}+b_{11} & a_{12} & a_{13} & a_{14} \\ a_{21}+b_{21} & a_{22} & a_{23} & a_{24} \\ a_{31}+b_{31} & b_{32} & a_{33} & a_{34} \\ a_{41}+b_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$
$$= \det \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} + \det \begin{bmatrix} b_{11} & a_{12} & a_{13} & a_{14} \\ b_{21} & a_{22} & a_{23} & a_{24} \\ b_{31} & a_{32} & a_{33} & a_{34} \\ b_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix},$$

four times, we can expand $\det \Lambda(\rho)$ into four determinants, such that

$$\det \Lambda(\rho) = \det \Lambda_0(\rho) + \sum_{j=1}^{3} \det \Lambda_j(\rho).$$

In each $\Lambda_j(\rho)$, $j = 1, 2, 3$, there is at least one column having $1/\rho$ as a common factor. Hence we get (3.18).

The other cases (3.18)′, (3.19), and (3.19)′ can all be treated in the same manner. So the proof is complete. □

Proof of Lemma 3.2. We proceed in two steps.

(1) There exists $k_1 \in \mathbb{Z}^+$ such that for $k \in \mathbb{Z}$, $|k| \geqq k_1$, there exists one and only one solution $\tilde{\rho}$ of (3.9) in G_{ik}, for $i = 1, 2$.

We need only consider the case $i = 1$, $k \geqq k_1$. From Lemma 3.1, equation (3.9) can be written as

$$4b(0)b(\pi)K_1K_2[E_1(\rho) - E_4(\rho)][E_3(\rho) - E_2(\rho)] + \rho^{-1}[g_1(\rho)E_3(\rho) + g_2(\rho)] = 0,$$

$$E_1(\rho) - E_4(\rho) + \frac{1}{\rho} \cdot \frac{g_1(\rho)E_3(\rho) + g_2(\rho)}{4b(0)b(\pi)K_1K_2[E_3(\rho) - E_2(\rho)]} = 0, \tag{5.1}$$

or

$$E_1(\rho) - E_4(\rho) + \frac{1}{\rho} g(\rho) = 0, \tag{5.2}$$

where $g(\rho)$ is uniformly bounded for $\rho \in G_{1k}^+$, $k \geqq k_1$. Define $F(\rho) = E_1(\rho) - E_4(\rho)$. For $\rho \in \partial G_{1k}^+$,

$$\begin{aligned} |F(\rho)| &= |e^{[\rho_1(k)+\varepsilon_1+i\varepsilon_2](\alpha+i\beta)} - e^{-[\rho_1(k)+\varepsilon_1+i\varepsilon_2](\alpha+i\beta)}| \\ &= |e^{(\varepsilon_1+i\varepsilon_2)(\alpha+i\beta)} - e^{-(\varepsilon_1+i\varepsilon_2)(\alpha+i\beta)}| \quad \text{(by (3.12))} \\ &= |e^{-(\varepsilon_1\alpha-\varepsilon_2\beta)} \cdot e^{-i(\varepsilon_1\beta+\varepsilon_2\alpha)}| \cdot |e^{2(\varepsilon_1\alpha-\varepsilon_2\beta)}\, e^{2i(\varepsilon_1\beta+\varepsilon_2\alpha)} - 1|. \end{aligned}$$

Refer to Fig. 3. On l_1: $\varepsilon_1 = 1$, $-1 \leqq \varepsilon_2 \leqq 1$, we have

$$\begin{aligned} |F(\rho)| &= |e^{-(\alpha-\varepsilon_2\beta)}| \cdot |e^{2(\alpha-\varepsilon_2\beta)}\, e^{2i(\beta+\varepsilon_2\alpha)} - 1| \\ &\geqq e^{-(\alpha-\beta)}[e^{2(\alpha-\beta)} - 1] \equiv M_1 > 0 \qquad (\because \alpha > \beta). \end{aligned}$$

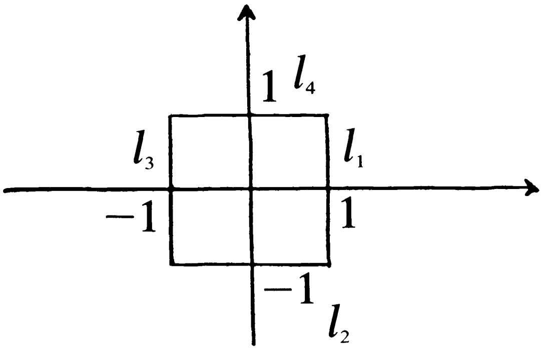


FIG. 3. *A small neighborhood.*

On l_3: $\varepsilon_1=-1, -1\leqq\varepsilon_2\leqq 1$,

$$\begin{aligned}|F(\rho)|&=|e^{-(-\alpha-\varepsilon_2\beta)}|\cdot|e^{2(-\alpha-\varepsilon_2\beta)}\,e^{2i(-\beta+\varepsilon_2\alpha)}-1|\\&\geqq e^{(\alpha-\beta)}[1-e^{-2(\alpha-\beta)}]\equiv M_2>0.\end{aligned}$$

On l_2: $-1\leqq\varepsilon_1\leqq 1$, $\varepsilon_2=-1$,

$$\begin{aligned}|F(\rho)|&=|e^{-(\varepsilon_1\alpha+\beta)}|\cdot|e^{2(\varepsilon_1\alpha+\beta)}\cdot e^{2i(\varepsilon_1\beta-\alpha)}-1|\\&\geqq e^{-(\alpha+\beta)}\cdot|e^{2(\varepsilon_1\alpha+\beta)}\,e^{2i(\varepsilon_1\beta-\alpha)}-1|.\end{aligned}$$

We claim that

$$\min_{-1\leqq\varepsilon_1\leqq 1}|e^{2(\varepsilon_1\alpha+\beta)}\,e^{2i(\varepsilon_1\beta-\alpha)}-1|>0.$$

Otherwise, there exists $\varepsilon_1^0\in[-1,1]$ such that

$$e^{2(\varepsilon_1^0\alpha+\beta)}\cdot e^{2i(\varepsilon_1^0\beta-\alpha)}-1=0,$$

implying

$$\begin{matrix}2(\varepsilon_1^0\alpha+\beta)=0\\2(\varepsilon_1^0\beta-\alpha)=0\end{matrix},\quad\text{i.e.,}\quad-\frac{\beta}{\alpha}=\varepsilon_1^0=\frac{\alpha}{\beta}.$$

So $\alpha^2+\beta^2=0$, $\alpha=\beta=0$. This is a contradiction. Therefore on l_2,

$$|F(\rho)|\geqq M_3>0\quad\text{for some } M_3.$$

Finally, on l_4: $-1\leqq\varepsilon_1\leqq 1$, $\varepsilon_2=1$,

$$\begin{aligned}|F(\rho)|&=|e^{-(\varepsilon_1\alpha-\beta)}|\cdot|e^{2(\varepsilon_1\alpha-\beta)}\cdot e^{2i(\varepsilon_b\beta+\alpha)}-1|\\&\geqq e^{-(\alpha-\beta)}\cdot|e^{2(\varepsilon_1\alpha-\beta)}\,e^{2i(\varepsilon_i\beta+\alpha)}-1|\geqq M_4>0,\end{aligned}$$

for some $M_4>0$, for the same reason as on l_3.

Let $M_5\equiv\min(M_1, M_2, M_3, M_4)$.

The uniform boundedness of g means that there exists $M_6>0$ such that

$$|g(p)|\leqq M_6\quad\text{for all } \rho\in G_{1k}^+,\ k\geqq k_0.$$

Choose $k_1\geqq k_0$ such that for $\rho\in G_{1k}^+$,

$$|\rho|>\frac{2M_6}{M_5}.$$

Then for $\rho \in \partial G^{1k}$, $k \geqq k_1$, we have

$$|E_1(\rho) - E_4(\rho)| \geqq M_5 > \frac{M_5}{2} \geqq \left| \frac{1}{\rho} g(\rho) \right|.$$

Since $E_1(\rho)$, $E_4(\rho)$, and $\rho^{-1}g(\rho)$ are all analytic in ρ, by Rouché's theorem there exists one and only one solution $\tilde{\rho}$ of (3.7) in G^+_{1k}. The other cases can be treated similarly.

(2) For $|\rho|$ large, there are no solutions $\tilde{\rho}$ of (3.9) outside $G_1 \cup G_2$.

From (5.1) and (5.2), we can write

$$e^{2\tilde{\rho}\Gamma_1(\pi)} - 1 + \frac{1}{\tilde{\rho}} \frac{g(\tilde{\rho})}{e^{-\tilde{\rho}\Gamma_1(\pi)}} = 0$$

or

$$e^{2\tilde{\rho}\Gamma_1(\pi)} - 1 + \frac{1}{\rho} \hat{g}(\tilde{\rho}) = 0. \tag{5.3}$$

Assume that as $|\tilde{\rho}| \to \infty$, $\exp[2\tilde{\rho}\Gamma_1(\pi)]$ becomes unbounded. Then from our foregoing discussions, we have

$$|\hat{g}(\tilde{\rho})| \text{ bounded as } |\tilde{\rho}| \to \infty, \qquad |\exp[2\tilde{\rho}\Gamma_1(\pi)]| \to \infty.$$

Therefore from (5.3)

$$\begin{aligned} \tilde{\rho} &= \frac{1}{2\Gamma_1(\pi)} \ln\left(1 - \frac{\hat{g}(\rho)}{\rho}\right) + i\frac{k\pi}{\Gamma_1(\pi)} \\ &= \rho_1(k) + \frac{1}{2\Gamma_1(\pi)} \ln\left(1 - \frac{\hat{g}(\rho)}{\rho}\right), \end{aligned}$$

showing that $\tilde{\rho} \in G_1$ for $|\tilde{\rho}|$ large.

If instead we use equations derived from (3.19), (3.19)′, then $\tilde{\rho} \in G_2$ for $|\rho|$ large. □

REFERENCES

[1] G. D. Birkhoff, *On the asymptotic character of the solutions of certain linear differential equations containing a parameter*, Trans. Amer. Math. Soc., 9 (1908), pp. 219-231.

[2] ———, *Boundary value and expansion problems of ordinary differential equations*, Trans. Amer. Math. Soc., 9 (1908), pp. 373-395.

[3] G. D. Birkhoff and R. E. Langer, *The boundary problems and developments associated with a system of ordinary linear differential equations of the first order*, Proc. Amer. Acad. Arts Sci., (2) 58 (1923), pp. 51-128.

[4] G. Chen, S. G. Krantz, D. W. Ma, C. E. Wayne, and H. H. West, *The Euler-Bernoulli beam equation with boundary energy dissipation*, in Operator Methods for Optimal Control Problems, S. J. Lee ed., Marcel Dekker, New York, 1988, pp. 67-96.

[5] G. Chen and D. L. Russell, *A mathematical model for linear elastic systems with structural damping*, Quart. Appl. Math., 39 (1982), pp. 433-454.

[6] G. Chen, S. G. Krantz, D. L. Russell, C. E. Wayne, H. H. West, and J. Zhou, *Modelling, analysis and testing of dissipative beam joints-experiments and data smoothing*, Mathl. Comput. Modelling, 11, Pergamon Press, New York, 1988, pp. 1011-1016.

[7] G. Chen, S. G. Krantz, D. L. Russell, C. E. Wayne, H. H. West, and M. P. Coleman, *Analysis, designs and behavior of dissipative joints for coupled beams*, SIAM J. Appl. Math., 49 (1989), pp. 1665-1693.

[8] G. Chen, S. A. Fulling, F. J. Narcowich, and C. Qi, *An asymptotic average decay rate for the wave equation with variable coefficient viscous damping*, SIAM J. Appl. Math., 50 (1990), pp. 1341-1347.

[9] G. CHEN AND J. ZHOU, *The wave propagation method for the analysis of boundary stabilization in vibrating structures*, SIAM J. Appl. Math., 50 (1990), pp. 1254-1283.

[10] F. L. HUANG, *On the holomorphic property of the semigroup associated with linear elastic systems with structural damping*, Acta Math. Sci., 5 (1985), pp. 271-277.

[11] D. L. RUSSELL, *On mathematical models for the elastic beam with frequency proportional damping*, to appear.

[12] ———, *On the positive square root of the fourth derivative operator*, Quart. Appl. Math. 46 (1966), pp. 751-773.

[13] H. K. WANG, *Asymptotic analysis of eigenfrequencies of Euler-Bernoulli beam equation with structural damping*, in Proc. 28th Annual IEEE Conference on Decision and Control, Tampa, FL, IEEE Computer Society, Washington, DC, 1989, pp. 2042-2044.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 368-380, March 1991

EXPONENTIAL DECAY OF THE ENERGY OF A ONE-DIMENSIONAL NONHOMOGENEOUS MEDIUM*

JONG UHN KIM†

Abstract. The energy method based on wave propagation is used to show that locally distributed damping can stabilize the motion of a composite bar exponentially fast.

Key words. exponential decay, nonhomogeneous bar, energy method

AMS(MOS) subject classifications. 35R05, 73K05, 93D20

Introduction. In this paper we prove that locally distributed damping can stabilize the motion of a one-dimensional nonhomogeneous medium exponentially fast. Specifically, we consider either longitudinal or torsional vibration of a nonhomogeneous bar. The bar is composed of two different segments, only one of which is damped.

The model equations are

$$m_1(x)\frac{\partial^2 u}{\partial t^2}-\frac{\partial}{\partial x}\left(a_1(x)\frac{\partial u}{\partial x}\right)=0 \quad \text{in } (0,\infty)\times(L_1, L_2), \tag{0.1}$$

$$m_2(x)\frac{\partial^2 u}{\partial t^2}-\frac{\partial}{\partial x}\left(a_2(x)\frac{\partial u}{\partial x}\right)+b(x)\frac{\partial u}{\partial t}=0 \quad \text{in } (0,\infty)\times(L_2, L_3), \tag{0.2}$$

where $L_1 < L_2 < L_3$.

The bar is represented by the interval (L_1, L_3) and consists of two different parts, represented by (L_1, L_2) and (L_2, L_3), respectively. Here, $m_i(x)$, $a_i(x)$, $i = 1, 2$, and $b(x)$ are appropriate physical coefficients accociated with each segment. At the interface $x = L_2$, we impose the transmission condition:

$$\lim_{x\to L_2-} u(t,x) = \lim_{x\to L_2+} u(t,x), \tag{0.3}$$

$$\lim_{x\to L_2-} a_1(x)\frac{\partial u}{\partial x}(t,x) = \lim_{x\to L_2+} a_2(x)\frac{\partial u}{\partial x}(t,x). \tag{0.4}$$

At the boundary $x = L_1, L_3$, we assume

$$u(t, L_1) = u(t, L_3) = 0. \tag{0.5}$$

It is known that the energy of a single bar with damping distributed over the whole interval decays exponentially fast with the homogeneous Dirichlet boundary condition (see [1]). Here we join two different bars, only one of which has damping. An obvious question is whether or not the energy of the combined system decays exponentially fast. Similar problems have been discussed in [2], [3], and [7]. Chen gave a talk on [3] at VPI and his talk motivated the present work. Reference [3] presents a general theorem on an abstract evolution equation with partially distributed damping. The result of [3] can cover various types of evolution equations. The conditions in [3] for

* Received by the editors April 17, 1989; accepted for publication (in revised form) April 27, 1990.

† Department of Mathematics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061. This research was supported by Air Force Office of Scientific Research grants AFOSR-86-0085 and AFOSR-89-0268.

the exponential decay of energy involve eigenvalues and eigenfunctions of the associated stationary system. Chen et al. also discussed some specific examples for which these conditions can be easily verified. In fact, if $m_1 = m_2 = \text{const.}$, $a_1 = a_2 = \text{const.}$, and $b(x) > 0$ on a certain subinterval, the above problem reduces to one of the examples treated in [3]. However, when m_i and a_i, $i = 1, 2$, are variable, it is not obvious whether we can verify all the conditions needed for their theorem. The well-established properties of eigenvalues and eigenfunctions of the Sturm-Liouville problem are not sufficient. In this paper, we employ the energy method in the same spirit as in [7]. Rauch and Taylor [7] treated problems on compact manifolds of any dimension without boundary and also the case of a one-dimensional bounded domain with boundary. Their discussion of the latter case gives a clue for our problem. But more involved energy estimates are necessary since the operator (after our problem is put in the framework of [7]) is not strictly dissipative (according to [7]) at any point of the interval. We also need a quantitative analysis of the energy transmission across the interface. The main tool we shall use is a multiplier technique due to Lasiecka, Lions, and Triggiani [5]. Our argument depends heavily on the phenomenon of wave propagation and does not extend to nonhyperbolic evolution equations such as Euler's beam equation.

In § 1, we present some technical preliminaries, and in § 2, we state the main result and give the proof.

Notation. $(\ )_x$ or $[\ ]_x$ denotes an interval for the x variable. Similarly, $(\ )_t$ or $[\ ]_t$ is an interval for the t variable. For the derivatives of a function, we use the following notation:

$$\partial_t f = f_t = \frac{\partial f}{\partial t}, \quad \partial_x f = f_x = \frac{\partial f}{\partial x}, \quad \partial_{xx} f = f_{xx} = \frac{\partial^2 f}{\partial x^2}, \text{ etc.}$$

When Ω is an open subset of R^n, $n \geqq 1$, then $\mathscr{D}'(\Omega)$ is the space of distributions, and $H^m(\Omega)$ and $H_0^m(\Omega)$ are the standard notation for the Sobolev spaces.

When E is a Banach space and I is an open interval, we use the following notation:

(0.6) $L^p(I; E) =$ the set of all E-valued strongly measurable L^p functions on I.

(0.7) $C_0^m(I; E) = \{f \in C^m(I; E)\colon \operatorname{supp} f \text{ is a compact subset of } I\}$.

(0.8) $\rho_\varepsilon(t) = \rho(t/\varepsilon)/\varepsilon$, where $\rho(t) =$ a nonnegative even C^∞ function with support in $(-1, 1)$ and integral one.

1. Preliminaries.

We shall present some properties of one-dimensional hyperbolic equations, which will be used later.

LEMMA 1.1. *Let $A(x) \in C^1([J_1, J_2])$ and $B(x) \in C([J_1, J_2])$ be 2×2 matrices. Suppose that $A(x)$ has real eigenvalues $\lambda_1(x)$ and $\lambda_2(x)$ such that $0 < \alpha \leqq \lambda_1(x) \leqq \beta$ and $-\beta \leqq \lambda_2(x) \leqq -\alpha$, for all $x \in [J_1, J_2]$, for some positive constants $\alpha < \beta$. Then, for given $h(t) \in L^2(T_1, T_2)$, there is a function $v \in C([J_1, J_2]_x : L^2(T_1, T_2)_t)$ satisfying the hyperbolic system*

$$\partial_t v - A(x)\,\partial_x v + B(x) v = 0 \quad \text{in } (T_1, T_2) \times (J_1, J_2), \tag{1.1}$$

$$v(t, J_2) = h(t) \quad \text{for almost all } t \in (T_1, T_2). \tag{1.2}$$

Furthermore, for $0 < c < \min(J_2 - J_1, \alpha(T_2 - T_1)/2)$, v is uniquely determined in a trapezoidal region $\sum = \{(t, x)\colon J_2 - c \leqq x \leqq J_2,\ T_1 + (J_2 - x)/\alpha \leqq t \leqq T_2 - (J_2 - x)/\alpha\}$, and

$$\int_{T_1 + c/\alpha}^{T_2 - c/\alpha} |v(t, x)|^2\, dt \leqq M \int_{T_1}^{T_2} |h(t)|^2\, dt \tag{1.3}$$

holds for each $J_2 - c \leqq x \leqq J_2$, where M is a positive constant independent of x, v, and h.

Proof. The proof seems to be known, hence we shall only sketch it. First, extend $h(t)$ such that $\tilde{h}(t)=h(t)$, for $T_1 \leqq t \leqq T_2$ and $\tilde{h}(t)=0$, otherwise. Next, choose a sequence $\{h_n(t)\} \subset C_0^\infty((-\infty,\infty))$ which approximates $\tilde{h}(t)$ in $L^2(-\infty,\infty)$. For each $h_n(t)$, we use the method of characteristics to find $v_n \in C^1((-\infty,\infty)_t \times [J_1, J_2]_x)$ satisfying (1.1) and $v_n(t, J_2)=h_n(t)$, for all $t \in (-\infty,\infty)$. We then obtain a standard energy estimate for v_n in terms of the L^2 norm of $h_n(t)$. By means of this estimate, we can easily establish the existence of $v \in C([J_1, J_2]_x; L^2(T_1, T_2))$ satisfying (1.1)-(1.3). The uniqueness of v in $\sum$ is well known and the proof will be omitted.

LEMMA 1.2. *Let $\Omega=(T_1, T_2)_t \times (J_1, J_2)_x$ and assume that $b(x) \in C([J_1, J_2])$, $m(x)$ and $a(x) \in C^1([J_1, J_2])$ with $m(x), a(x) \geqq c > 0$, for all $x \in [J_1, J_2]$, for some constant c. If $u \in H^1(\Omega)$ satisfies*

$$(1.4) \qquad m(x)\,\partial_{tt}u - \partial_x(a(x)\,\partial_x u) + b(x)\,\partial_t u = 0 \quad \text{in } \mathscr{D}'(\Omega),$$

then for any small $\delta > 0$, we have

$$(1.5) \qquad \partial_t u, \partial_x u \in C([J_1, J_2]_x; L^2(T_1+\delta, T_2-\delta)_t),$$

$$(1.6) \qquad \partial_t u, \partial_x u \in C([T_1, T_2]_t; L^2(J_1+\delta, J_2-\delta)_x)$$

and

$$(1.7) \qquad \begin{aligned} &\|\partial_t u(x)\|_{L^2(T_1+\delta, T_2-\delta)_t} + \|\partial_x u(x)\|_{L^2(T_1+\delta, T_2-\delta)_t} \\ &\quad \leqq M_\delta(\|\partial_t u\|_{L^2(\Omega)} + \|\partial_x u\|_{L^2(\Omega)}) \quad \text{for each } x \in [J_1, J_2], \end{aligned}$$

$$(1.8) \qquad \begin{aligned} &\|\partial_t u(t)\|_{L^2(J_1+\delta, J_2-\delta)_x} + \|\partial_x u(t)\|_{L^2(J_1+\delta, J_2-\delta)_x} \\ &\quad \leqq M_\delta(\|\partial_t u\|_{L^2(\Omega)} + \|\partial_x u\|_{L^2(\Omega)}) \quad \text{for each } t \in [T_1, T_2], \end{aligned}$$

where M_δ stands for positive constants depending only on δ.

Proof. Fix any small $\delta > 0$ and $\xi > 0$. Choose $\varphi_1(t) \in C_0^\infty((T_1, T_2))$ and $\varphi_2(x) \in C_0^\infty((J_1, J_2))$ such that $\varphi_1(t)=1$, for $t \in [T_1+\delta, T_2-\delta]$ and $\varphi_2(x)=1$, for $x \in [J_1+\xi, J_2-\xi]$.

Next we set $\varphi(t,x)=\varphi_1(t)\varphi_2(x)$ and $v=\varphi u$. Then v satisfies

$$(1.9) \qquad m(x)\,\partial_{tt}v - \partial_x(a(x)\,\partial_x v) = g \quad \text{in } \mathscr{D}'(\Omega),$$

where

$$g = -b(x)\varphi u_t + 2m\varphi_t u_t + m\varphi_{tt}u - 2a\varphi_x u_x - a_x\varphi_x u - au\varphi_{xx}.$$

We let

$$(1.10) \qquad v_\varepsilon = v * \rho_\varepsilon, \quad g_\varepsilon = g * \rho_\varepsilon \quad \text{for small } \varepsilon > 0,$$

where $\rho_\varepsilon(t)$ is the regularizer defined by (0.8) and the convolution is taken in the t variable. Then, for every small $\varepsilon > 0$,

$$v_\varepsilon \in C_0^\infty((T_1, T_2)_t; H_0^1(J_1, J_2)_x)$$

and

$$(1.11) \qquad m(x)\,\partial_{tt}v_\varepsilon - \partial_x(a(x)\,\partial_x v_\varepsilon) = g_\varepsilon$$

holds. Hence, $\partial_x v_\varepsilon \in C_0^\infty((T_1, T_2)_t; H_0^1(J_1, J_2)_x)$. We shall show that $\{\partial_t v_\varepsilon(x)\}$ and $\{\partial_x v_\varepsilon(x)\}$ converge in $L^2(T_1, T_2)_t$ as $\varepsilon \to 0$ uniformly in $x \in [J_1, J_2]$. Let $w = v_\varepsilon - v_{\varepsilon'}$. Then

$$(1.12) \qquad m(x)\,\partial_{tt}w - \partial_x(a(x)\,\partial_x w) = g_\varepsilon - g_{\varepsilon'}.$$

Let us multiply (1.12) by w_x, choose any $x_0 \in [J_1, J_2]$, and integrate over $[T_1, T_2] \times [x_0, J_2]$:

$$\begin{aligned}(1.13)\qquad &\int_{T_1}^{T_2} m(x_0) w_t(t, x_0)^2\, dt + \int_{T_1}^{T_2} a(x_0) w_x(t, x_0)^2\, dt \\ &\leqq M \int_{T_1}^{T_2} \int_{J_1}^{J_2} (w_t^2 + w_x^2)\, dx\, dt + \int_{T_1}^{T_2} \int_{J_1}^{J_2} (g_\varepsilon - g_{\varepsilon'})^2\, dx\, dt,\end{aligned}$$

where M is a positive constant independent of x_0, ε, and ε'. We next choose any $t_0 \in [T_1, T_2]$, multiply (1.12) by w_t, and integrate over $[T_1, t_0] \times [J_1, J_2]$:

$$(1.14)\quad \int_{J_1}^{J_2} m(x) w_t(t_0, x)^2\, dx + \int_{J_1}^{J_2} a(x) w_x(t_0, x)^2\, dx = 2 \int_{T_1}^{t_0} \int_{J_1}^{J_2} (g_\varepsilon - g_{\varepsilon'}) w_t\, dx\, dt.$$

By virtue of Gronwall's inequality, it follows that

$$\begin{aligned}(1.15)\qquad &\int_{J_1}^{J_2} m(x) w_t(t, x)^2\, dx + \int_{J_1}^{J_2} a(x) w_x(t, x)^2\, dx \\ &\leqq M \int_{T_1}^{T_2} \int_{J_1}^{J_2} (g_\varepsilon - g_{\varepsilon'})^2\, dx\, dt \quad \text{for all } t \in [T_1, T_2],\end{aligned}$$

where M is a positive constant independent of ε and ε'. Combining (1.13) and (1.15), we obtain

$$(1.16)\qquad \int_{T_1}^{T_2} w_t(t, x_0)^2\, dt + \int_{T_1}^{T_2} w_x(t, x_0)^2\, dt \leqq M \int_{T_1}^{T_2} \int_{J_1}^{J_2} (g_\varepsilon - g_{\varepsilon'})^2\, dx\, dt,$$

where M is a positive constant independent of x_0, ε, and ε'. It follows from (1.16) that $\{\partial_t v_\varepsilon(x)\}$ and $\{\partial_x v_\varepsilon(x)\}$ converge in $L^2(T_1, T_2)_t$ uniformly in $x \in [J_1, J_2]$. Therefore, $\partial_t u, \partial_x u \in C([J_1 + \xi, J_2 - \xi]_x; L^2(T_1 + \delta, T_2 - \delta)_t)$. Since ξ is arbitrary, we conclude that

$$(1.17)\qquad \partial_t u, \partial_x u \in C((J_1, J_2)_x; L^2(T_1 + \delta, T_2 - \delta)_t),$$

for each $0 < \delta < (T_2 - T_1)/2$. Next we extend $m(x)$, $a(x)$, and $b(x)$ such that $\tilde{b}(x) \in C([J_1 - 1, J_2 + 1])$, $\tilde{m}(x)$ and $\tilde{a}(x) \in C^1([J_1 - 1, J_2 + 1])$ with $\tilde{m}(x), \tilde{a}(x) \geqq c/2 > 0$ for all $x \in [J_1 - 1, J_2 + 1]$. Then, we consider the hyperbolic system

$$(1.18)\quad \partial_t U - \begin{pmatrix} 0, & 1 \\ \tilde{a}(x)/\tilde{m}(x), & 0 \end{pmatrix} \partial_x U + \begin{pmatrix} 0, & 0 \\ -\tilde{a}_x(x)/\tilde{m}(x), & \tilde{b}(x)/\tilde{m}(x) \end{pmatrix} U = 0.$$

Let α be a positive constant such that

$$(1.19)\qquad \tilde{a}(x)/\tilde{m}(x) \geqq \alpha^2 \quad \text{for all } x \in [J_1 - 1, J_2 + 1].$$

Fix any small $\eta > 0$. We apply Lemma 1.1 by choosing $\binom{\partial_x u}{\partial_t u}$ at $x = J_1 + \eta$ as the initial datum to find a function

$$U \in C([J_1 - 1, J_1 + \eta]_x; L^2(T_1 + \delta, T_2 - \delta)_t),$$

which is equal to $\binom{\partial_x u}{\partial_t u}$ in $(J_1, J_1 + \eta) \times (T_1 + \delta + \eta/\alpha, T_2 - \delta - \eta/\alpha)$. Since δ and η are arbitrary, we derive that $\partial_t u, \partial_x u \in C([J_1, J_2)_x; L^2(T_1 + \delta, T_2 - \delta)_t)$ for each small $\delta > 0$. Similarly, we can apply Lemma 1.1 near $x = J_2$ and obtain the continuity at $x = J_2$. Now the proof of (1.5) is complete.

To obtain the estimate (1.7), we first derive an analogue of (1.13) for v_ε defined by (1.10):

$$(1.20)\quad \int_{T_1}^{T_2} v_{\varepsilon t}(t,x)^2\,dt + \int_{T_1}^{T_2} v_{\varepsilon x}(t,x)^2\,dt \leqq M(\delta,\xi)\|u\|^2_{H^1(\Omega)} \quad \text{for all } x\in[J_1,J_2],$$

where $M(\delta,\xi)$ is a positive constant depending only on δ and ξ. It follows from (1.20) that

$$(1.21)\qquad \int_{T_1+\delta}^{T_2-\delta} (u_t(t,x)^2+u_x(t,x)^2)\,dt \leqq M(\delta,\xi)\|u\|^2_{H^1(\Omega)},$$

for each $x\in[J_1+\xi, J_2-\xi]$.

Again we apply Lemma 1.1 with the initial datum $\binom{\partial_x u}{\partial_t u}$ at $x=J_1+\xi$ and at $x=J_2-\xi$, respectively. Formulas (1.3) and (1.21) imply

$$(1.22)\qquad \begin{aligned}&\int_{T_1+\delta+\xi/\alpha}^{T_2-\delta-\xi/\alpha} (u_t(t,x)^2+u_x(t,x)^2)\,dt\\ &\leqq M(\delta,\xi)\|u\|^2_{H^1(\Omega)} \quad \text{for all } x\in[J_1,J_1+\xi]\cup[J_2-\xi,J_2],\end{aligned}$$

where $M(\delta,\xi)$ is a positive constant depending only on δ and ξ. Since δ and ξ are arbitrary, it follows from (1.21) and (1.22) that for each $\delta>0$,

$$(1.23)\qquad \int_{T_1+\delta}^{T_2-\delta} (u_t(t,x)^2+u_x(t,x)^2)\,dt \leqq M_\delta\|u\|^2_{H^1(\Omega)},$$

for all $x\in[J_1,J_2]$. Next we let

$$(1.24)\qquad \tilde u = u - \frac{1}{\operatorname{meas}(\Omega)}\int_\Omega u(t,x)\,dt\,dx.$$

Then, $\tilde u\in H^1(\Omega)$ and satisfies (1.4). Therefore, $\tilde u$ satisfies (1.23)

$$(1.25)\qquad \int_{T_1+\delta}^{T_2-\delta} (\tilde u_t(t,x)^2+\tilde u_x(t,x)^2)\,dt \leqq M_\delta\|\tilde u\|^2_{H^1(\Omega)},$$

for all $x\in[J_1,J_2]$. But, $\partial_t u=\partial_t\tilde u$, $\partial_x u=\partial_x\tilde u$, and, by Poincare's inequality,

$$(1.26)\qquad \|\tilde u\|_{H^1(\Omega)} \leqq C(\|\partial_t\tilde u\|_{L^2(\Omega)}+\|\partial_x\tilde u\|_{L^2(\Omega)})$$

holds. Now (1.7) is obvious. The proof of (1.6) and (1.8) can be carried out in a similar manner, and we omit the details.

LEMMA 1.3. *Let $\Omega=(T_1,T_2)\times(J_1,J_2)$, $b(x)\in C([J_1,J_2])$, and $m(x), a(x)\in C^1([J_1,J_2])$ with $m(x), a(x)\geqq c>0$, for all $x\in[J_1,J_2]$. Suppose that u satisfies*

$$(1.27)\qquad u\in H^1(\Omega),$$

$$(1.28)\qquad m(x)\,\partial_{tt}u-\partial_x(a(x)\,\partial_x u)+b(x)\,\partial_t u=0 \quad \text{in } \mathscr{D}'(\Omega).$$

Then, it holds that

$$(1.29)\qquad \|\partial_x u\|_{L^2(\Omega_\delta)} \leqq M_\delta(\|\partial_t u\|_{L^2(\Omega)}+\|u\|_{L^2(\Omega)}),$$

for each small $\delta>0$, where

$$\Omega_\delta=(T_1+\delta, T_2-\delta)\times(J_1+\delta, J_2-\delta),$$

and M_δ is a positive constant independent of u and dependent only on δ.

Proof. It follows from (1.9)

$$(1.30)\quad -\int_\Omega m(x)v_t(t,x)^2\,dt\,dx+\int_\Omega a(x)v_x(t,x)^2\,dt\,dx=\int_\Omega g(t,x)v(t,x)\,dt\,dx.$$

Using $\varphi_x u_x=\partial_x(\varphi_x u)-\varphi_{xx}u$ and integration by parts, we can easily derive (1.29) from (1.30).

2. Main result. Throughout this section, we assume the following:

(i) $m_1(x), a_1(x)\in C^1([L_1,L_2])$ and

$$(2.1)\qquad m_1(x), a_1(x)\geqq\beta_1>0\quad\text{for all } x\in[L_1,L_2].$$

(ii) $m_2(x), a_2(x)\in C^1([L_2,L_3])$ and

$$(2.2)\qquad m_2(x), a_2(x)\geqq\beta_2>0\quad\text{for all } x\in[L_2,L_3].$$

(iii) $b(x)\in C([L_2,L_3])$ and

$$(2.3)\qquad b(x)\geqq\beta_3>0\quad\text{for all } x\in[L_2,L_3].$$

Here β_i's are positive constants, and we do not assume that $a_1(L_2)=a_2(L_2)$ or $m_1(L_2)=m_2(L_2)$.

We define

$$(2.4)\qquad \zeta=4(L_3-L_1)\left\{\max_{x\in[L_1,L_2]}(m_1(x)/a_1(x))^{1/2}+\max_{x\in[L_2,L_3]}(m_2(x)/a_2(x))^{1/2}\right\}.$$

Before we present the main result on the uniform stabilization, the existence and uniqueness of solutions should be made precise.

LEMMA 2.1. *If $u_0(x)\in H_0^1(L_1,L_3)$ and $u_1(x)\in L^2(L_1,L_3)$, then there is a unique function u such that*

$$(2.5)\qquad u\in C([0,\infty)_t;H_0^1(L_1,L_3)_x),\qquad \partial_t u\in C([0,\infty)_t;L^2(L_1,L_3)_x),$$

$$(2.6)\qquad u(0,x)=u_0(x),\quad \partial_t u(0,x)=u_1(x)\quad \textit{for almost all } x\in(L_1,L_3),$$

$$(2.7)\qquad \begin{aligned}&-\int_0^\infty\int_{L_1}^{L_2} m_1(x)u_t\phi_t\,dx\,dt-\int_0^\infty\int_{L_2}^{L_3} m_2(x)u_t\phi_t\,dx\,dt\\&+\int_0^\infty\int_{L_1}^{L_2} a_1(x)u_x\phi_x\,dx\,dt+\int_0^\infty\int_{L_2}^{L_3} a_2(x)u_x\phi_x\,dx\,dt\\&+\int_0^\infty\int_{L_2}^{L_3} b(x)u_t\phi\,dx\,dt=0\end{aligned}$$

holds for every $\phi\in C_0^\infty((0,\infty)\times(L_1,L_3))$.

Proof. This lemma seems to be known. We can prove this by following the procedure in the proof of Theorem 8.1 of [6, p. 265] and Theorem 8.2 of [6, p. 275]. The argument is due to [8]. We omit the details.

Next we show that the above solution satisfies the transmission condition at $x=L_2$.

LEMMA 2.2. *For any $T>0$, the above solution u satisfies:*

$$(2.8)\qquad \partial_t u, \partial_x u\in C([L_1,L_2]_x;L^2(0,T)_t),$$

$$(2.9)\qquad \partial_t u, \partial_x u\in C([L_2,L_3]_x;L^2(0,T)_t),$$

$$(2.10)\qquad \lim_{x\to L_2-}u(t,x)=\lim_{x\to L_2+}u(t,x)\quad \textit{in } C([0,T]_t),$$

$$(2.11)\qquad \lim_{x\to L_2-}a_1(x)\,\partial_x u(t,x)=\lim_{x\to L_2+}a_2(x)\,\partial_x u(t,x)\quad \textit{in } L^2(0,T)_t.$$

Proof. It is easy to see that the solution can be uniquely extended to $(-\infty, \infty)_t$ so that

$$u \in C((-\infty, \infty)_t; H_0^1(L_1, L_3)_x), \tag{2.12}$$

$$\partial_t u \in C((-\infty, \infty)_t; L^2(L_1, L_3)_x). \tag{2.13}$$

Now (2.8) and (2.9) follow from Lemma 1.2. Condition (2.12) implies (2.10).

To show (2.11), we fix any $T > 0$ and $\psi \in C_0^\infty((0, T))$.

Let us denote by $[,]$ the L^2 inner product in $L^2(0, T)_t$. Then, according to (2.7),

$$\begin{aligned} &-\int_{L_1}^{L_2} m_1(x)[u_t, \psi_t](x)\phi(x)\,dx - \int_{L_2}^{L_3} m_2(x)[u_t, \psi_t](x)\phi(x)\,dx \\ &+\int_{L_1}^{L_2} a_1(x)[u_x, \psi](x)\phi_x(x)\,dx + \int_{L_2}^{L_3} a_2(x)[u_x, \psi](x)\phi_x(x)\,dx \\ &+\int_{L_2}^{L_3} b(x)[u_t, \psi](x)\phi(x)\,dx = 0, \end{aligned} \tag{2.14}$$

holds for each $\phi \in C_0^\infty((L_1, L_3))$.

This implies, by choosing $\phi \in C_0^\infty((L_1, L_2))$ and $\phi \in C_0^\infty((L_2, L_3))$,

$$-m_1(x)[u_t, \psi_t](x) - \partial_x(a_1(x)[u_x, \psi](x)) = 0 \quad \text{in } \mathscr{D}'((L_1, L_2)_x), \tag{2.15}$$

and

$$-m_2(x)[u_t, \psi_t](x) - \partial_x(a_2(x)[u_x, \psi](x)) + b(x)[u_t, \psi](x) = 0 \quad \text{in } \mathscr{D}'((L_2, L_3)_x). \tag{2.16}$$

Since $[u_t, \psi_t](x) \in L^2(L_1, L_3)$, we infer from (2.15) and (2.16) that $\partial_x(a_1(x)[u_x, \psi](x)) \in L^2(L_1, L_2)$ and $\partial_x(a_2(x)[u_x, \psi](x)) \in L^2(L_2, L_3)$, and that (2.15) and (2.16) hold in $L^2(L_1, L_2)$ and $L^2(L_2, L_3)$, respectively. Hence, for each $\phi \in C_0^\infty((L_1, L_3))$,

$$\begin{aligned} &-\int_{L_1}^{L_2} m_1(x)[u_t, \psi_t](x)\phi(x)\,dx - \int_{L_2}^{L_3} m_2(x)[u_t, \psi_t](x)\phi(x)\,dx \\ &-\int_{L_1}^{L_2} \phi(x)\,\partial_x(a_1(x)[u_x, \psi](x))\,dx - \int_{L_2}^{L_3} \phi(x)\,\partial_x(a_2(x)[u_x, \psi](x))\,dx \\ &+\int_{L_2}^{L_3} b(x)[u_t, \psi](x)\phi(x)\,dx = 0. \end{aligned} \tag{2.17}$$

After integration by parts using (2.8) and (2.9), we compare (2.14) and (2.17) so that

$$\lim_{x \to L_2-} \phi(x)a_1(x)[u_x, \psi](x) = \lim_{x \to L_2+} \phi(x)a_2(x)[u_x, \psi](x). \tag{2.18}$$

Since (2.18) holds for all $\phi \in C_0^\infty((L_1, L_3))$ and $\psi \in C_0^\infty((0, T))$, (2.8) and (2.9) imply (2.11). This completes the proof of Lemma 2.2.

We are now ready to present the main result. Let u be the above solution and we set

$$E(t) = \langle u_t, u_t\rangle_1 + \langle u, u\rangle_2, \tag{2.19}$$

where $\langle,\rangle_1$ and $\langle,\rangle_2$ are defined by

$$\langle \varphi_1, \varphi_2\rangle_1 = \int_{L_1}^{L_2} m_1(x)\varphi_1(x)\varphi_2(x)\,dx + \int_{L_2}^{L_3} m_2(x)\varphi_1(x)\varphi_2(x)\,dx, \tag{2.20}$$

$$\langle \varphi_1, \varphi_2\rangle_2 = \int_{L_1}^{L_2} a_1(x)(\partial_x\varphi_1)(\partial_x\varphi_2)\,dx + \int_{L_2}^{L_3} a_2(x)(\partial_x\varphi_1)(\partial_x\varphi_2)\,dx. \tag{2.21}$$

The main result is Theorem 2.3.

THEOREM 2.3. *There are constants $C>0$ and $\alpha>0$ independent of u such that*

$$E(t)\leqq Ce^{-\alpha t}E(0) \quad \textit{for all } t\geqq 0. \tag{2.22}$$

The remainder of this section will be expended on the proof of (2.22).

The basic idea is to establish

$$E(T)\leqq k(T)E(0) \tag{2.23}$$

for some $T>0$, where $0<k(T)<1$ is a constant independent of u. Formula (2.23) implies (2.22). For (2.23), we need Lemma 2.4.

LEMMA 2.4. *Let $T>3\zeta$ (defined by (2.4)) and define a function space*

$$X=\{v\in H^1((0,T)\times(L_1,L_3)):\ v(t,L_1)=v(t,L_3)=0$$

$$\textit{for almost all } t\in(0,T), \textit{ and } v \textit{ satisfies}$$

$$(2.7)\ \textit{for every } \phi\in H_0^1((0,T)\times(L_1,L_3))\}.$$

Then for each $\varepsilon>0$, there is a positive constant $C(T,\varepsilon)$ such that

$$\int_\zeta^{T-\zeta}(\langle v_t,v_t\rangle_1+\langle v,v\rangle_2)\,dt\leqq C(T,\varepsilon)\int_0^T\int_{L_2}^{L_3}v_t^2\,dx\,dt+\varepsilon\int_0^T\int_{L_2}^{L_3}v^2\,dx\,dt, \tag{2.24}$$

for all $v\in X$.

We shall show that (2.24) implies (2.23).

First, $E(t)$ satisfies

$$E(t_2)=E(t_1)-2\int_{t_1}^{t_2}\int_{L_2}^{L_3}b(x)u_t^2\,dx\,dt \quad \text{for all } t_2\geqq t_1\geqq 0. \tag{2.25}$$

Since $u\in C([0,T]_t;H_0^1(L_1,L_3)_x)$ and $E(t)$ is nonincreasing in t by (2.25), it is easy to see that

$$\int_0^T\int_{L_2}^{L_3}u^2\,dx\,dt\leqq MTE(0), \tag{2.26}$$

where M is a positive constant independent of u and T. By choosing $\varepsilon=1/T$, it follows from (2.3), (2.24), (2.25), and (2.26) that

$$\int_\zeta^{T-\zeta}E(t)\,dt\leqq\frac{1}{2\beta_3}(E(0)-E(T))C\left(T,\frac{1}{T}\right)+ME(0). \tag{2.27}$$

Since $E(t)$ is nonincreasing in t, we have

$$TE(T)\leqq\int_\zeta^{T-\zeta}E(t)\,dt+2\zeta E(0), \tag{2.28}$$

which, combined with (2.27), yields

$$\left\{T+\frac{1}{2\beta_3}C\left(T,\frac{1}{T}\right)\right\}E(T)\leqq\left\{M+2\zeta+\frac{1}{2\beta_3}C\left(T,\frac{1}{T}\right)\right\}E(0). \tag{2.29}$$

Hence, by taking $T>M+2\zeta$, we get (2.23).

Before proceeding to prove Lemma 2.4, we need to show Lemma 2.5.

LEMMA 2.5. *Let $v\in X$. Then v satisfies*

$$\lim_{x\to L_2-}\partial_t v(t,x)=\lim_{x\to L_2+}\partial_t v(t,x) \quad \textit{in } L^2(\delta,T-\delta) \textit{ for any } \delta>0, \tag{2.30}$$

$$\lim_{x\to L_2-}a_1(x)\,\partial_x v(t,x)=\lim_{x\to L_2+}a_2(x)\,\partial_x v(t,x) \quad \textit{in } L^2(\delta,T-\delta) \textit{ for any } \delta>0. \tag{2.31}$$

Proof. Since $\partial_x v \in L^2((0, T) \times (L_1, L_3))$, we find that $v \in C([L_1, L_3]_x; L^2(0, T)_t)$ possibly after a modification on a set of measure zero. Thus, we have

$$\lim_{x \to L_2-} v(t, x) = \lim_{x \to L_2+} v(t, x) \quad \text{in } L^2(0, T)_t. \tag{2.32}$$

In the meantime, it follows from Lemma 1.2 that

$$\partial_t v, \partial_x v \in C([L_1, L_2]_x; L^2(\delta, T-\delta)_t), \tag{2.33}$$

$$\partial_t v, \partial_x v \in C([L_2, L_3]_x; L^2(\delta, T-\delta)_t), \tag{2.34}$$

for any $\delta > 0$, which, together with (2.32), yield (2.30). Equation (2.31) can be derived by the same argument as for (2.11), and we omit the details.

Proof of Lemma 2.4. Suppose that (2.24) is false. Then there is some $\varepsilon_0 > 0$ and sequences $\{C_n\}_{n=1}^{\infty}, \{v_n\}_{n=1}^{\infty} \subset X$ such that $C_n \to \infty$ and

$$\int_{\zeta}^{T-\zeta} (\langle \partial_t v_n, \partial_t v_n \rangle_1 + \langle v_n, v_n \rangle_2)\, dt > C_n \int_0^T \int_{L_2}^{L_3} (\partial_t v_n)^2\, dx\, dt + \varepsilon_0 \int_0^T \int_{L_2}^{L_3} v_n^2\, dx\, dt. \tag{2.35}$$

Meanwhile, it follows from Lemma 1.2 that for all $x \in [L_2, L_3]$,

$$\int_{3\zeta/4}^{T-3\zeta/4} \{(\partial_t v_n(t, x))^2 + (\partial_x v_n(t, x))^2\}\, dt \leqq C \int_{\zeta/2}^{T-\zeta/2} \int_{L_2}^{L_3} \{(\partial_t v_n)^2 + (\partial_x v_n)^2\}\, dx\, dt, \tag{2.36}$$

where C is a positive constant independent of v_n.

Noting that $U = \binom{\partial_x v_n}{\partial_t v_n}$ satisfies

$$\partial_t U - \begin{pmatrix} 0, & 1 \\ a_1(x)/m_1(x), & 0 \end{pmatrix} \partial_x U + \begin{pmatrix} 0, & 0 \\ -\partial_x a_1(x)/m_1(x), & 0 \end{pmatrix} U = 0, \tag{2.37}$$

in $(0, T) \times (L_1, L_2)$ and recalling (2.4), we can use (1.3) of Lemma 1.1 to derive for all $x \in [L_1, L_2]$,

$$\begin{aligned} &\int_{\zeta}^{T-\zeta} \{(\partial_t v_n(t, x))^2 + (\partial_x v_n(t, x))^2\}\, dt \\ &\qquad \leqq C \int_{3\zeta/4}^{T-3\zeta/4} \{(\partial_t v_n(t, L_2))^2 + (\partial_x v_n(t, L_2))^2\}\, dt, \end{aligned} \tag{2.38}$$

where C denotes a positive constant independent of v_n.

Now, we combine Lemma 2.5, (2.36), and (2.38) to find

$$\begin{aligned} &\int_{\zeta}^{T-\zeta} \int_{L_1}^{L_2} \{m_1(x)(\partial_t v_n)^2 + a_1(x)(\partial_x v_n)^2\}\, dx\, dt \\ &\qquad \leqq C \int_{\zeta/2}^{T-\zeta/2} \int_{L_2}^{L_3} \{m_2(x)(\partial_t v_n)^2 + a_2(x)(\partial_x v_n)^2\}\, dx\, dt, \end{aligned} \tag{2.39}$$

where C is a positive constant independent of v_n. Hence, it follows that

$$\begin{aligned} &\int_{\zeta/2}^{T-\zeta/2} \int_{L_2}^{L_3} \{m_2(x)(\partial_t v_n)^2 + a_2(x)(\partial_x v_n)^2\}\, dx\, dt \\ &\qquad > \frac{C_n}{C+1} \int_0^T \int_{L_2}^{L_3} (\partial_t v_n)^2\, dx\, dt + \frac{\varepsilon_0}{C+1} \int_0^T \int_{L_2}^{L_3} v_n^2\, dx\, dt. \end{aligned} \tag{2.40}$$

Now we set

$$w_n = v_n \Bigg/ \left(\int_{\zeta/2}^{T-\zeta/2} \int_{L_2}^{L_3} \{m_2(x)(\partial_t v_n)^2 + a_2(x)(\partial_x v_n)^2\}\, dx\, dt \right)^{1/2}. \tag{2.41}$$

Then $w_n \in X$ and

$$\int_{\zeta/2}^{T-\zeta/2}\int_{L_2}^{L_3}\{m_2(x)(\partial_t w_n)^2 + a_2(x)(\partial_x w_n)^2\}\,dx\,dt = 1, \tag{2.42}$$

$$1 > \frac{C_n}{C+1}\int_0^T\int_{L_2}^{L_3}(\partial_t w_n)^2\,dx\,dt + \frac{\varepsilon_0}{C+1}\int_0^T\int_{L_2}^{L_3} w_n^2\,dx\,dt, \tag{2.43}$$

hold for every n.

According to Lemma 1.3, we derive from (2.43)

$$\int_{\zeta/8}^{T-\zeta/8}\int_{L_2+\eta}^{L_3-\eta}(\partial_x w_n)^2\,dx\,dt \leqq M \quad \text{for all } n, \tag{2.44}$$

where $\eta = \frac{1}{4}(L_3 - L_2)$ and M is a positive constant. By means of (2.43), (2.44), we can apply Lemmas 1.1 and 1.2 in the same manner as for (2.39) to obtain

$$\int_{\zeta/4}^{T-\zeta/4}\left(\int_{L_2}^{L_2+\eta} + \int_{L_3-\eta}^{L_3}\right)(\partial_x w_n)^2\,dx\,dt \leqq M \quad \text{for all } n, \tag{2.45}$$

so that

$$\int_{\zeta/4}^{T-\zeta/4}\int_{L_2}^{L_3}\{(\partial_x w_n)^2 + (\partial_t w_n)^2\}\,dx\,dt \leqq M \quad \text{for all } n. \tag{2.46}$$

Again by the same argument as for (2.39), we obtain from (2.46)

$$\int_{\zeta/2}^{T-\zeta/2}\int_{L_1}^{L_2}\{(\partial_x w_n)^2 + (\partial_t w_n)^2\}\,dx\,dt \leqq M \quad \text{for all } n. \tag{2.47}$$

By (2.46) and (2.47), we can extract a subsequence still denoted by $\{w_n\}$ such that

$$\partial_x w_n \to \partial_x w_\infty \text{ weakly in } L^2((\zeta/2, T-\zeta/2)\times(L_1, L_3)) \cap L^2((\zeta/4, T-\zeta/4)\times(L_2, L_3)), \tag{2.48}$$

$$\partial_t w_n \to \partial_t w_\infty \text{ weakly in } L^2((\zeta/2, T-\zeta/2)\times(L_1, L_3)) \cap L^2((\zeta/4, T-\zeta/4)\times(L_2, L_3)), \tag{2.49}$$

for some $w_\infty \in L^2(\zeta/2, T-\zeta/2; H_0^1(L_1, L_3)_x) \cap H^1((\zeta/4, T-\zeta/4)\times(L_2, L_3))$ with $\partial_t w_\infty \in L^2((\zeta/2, T-\zeta/2)\times(L_1, L_3)) \cap L^2((\zeta/4, T-\zeta/4)\times(L_2, L_3))$. This w_∞ also satisfies

$$\int_{\zeta/2}^{T-\zeta/2}(-\langle\partial_t w_\infty, \partial_t\phi\rangle_1 + \langle w_\infty, \phi\rangle_2)\,dt + \int_{\zeta/2}^{T-\zeta/2}\int_{L_2}^{L_3} b(x)(\partial_t w_\infty)\phi\,dx\,dt = 0, \tag{2.50}$$

for every $\phi \in H_0^1((\zeta/2, T-\zeta/2)\times(L_1, L_3))$, and a similar equation for every $\phi \in H_0^1((\zeta/4, T-\zeta/4)\times(L_2, L_3))$. Next we shall investigate the implication of (2.42) and (2.43) on w_∞. The following version of the div-curl lemma is useful.

LEMMA 2.6. *Let $\{p_n\}_{n=1}^\infty$, $\{q_n\}_{n=1}^\infty$, $\{r_n\}_{n=1}^\infty$, and $\{s_n\}_{n=1}^\infty$ be bounded sequences in $L^2(\Omega)$, where Ω is a bounded open subset of R^2. Suppose that $\partial_x p_n = \partial_t q_n$, for each n and that $\{\partial_t r_n + \partial_x s_n\}_{n=1}^\infty$ is bounded in $L^2(\Omega)$. If $p_n \to p$, $q_n \to q$, $r_n \to r$ and $s_n \to s$ weakly in $L^2(\Omega)$, then*

$$p_n r_n + q_n s_n \to pr + qs \quad \text{in } \mathscr{D}'(\Omega).$$

Proof. See [4] for the proof.

By setting $p_n=\partial_t w_n$, $q_n=\partial_x w_n$, $r_n=m_2(x)\,\partial_t w_n$, $s_n=-a_2(x)\,\partial_x w_n$, we use (2.46), (2.48), (2.49), and

$$m_2(x)\,\partial_{tt}w_n-\partial_x(a_2(x)\,\partial_x w_n)+b(x)\,\partial_t w_n=0 \quad \text{in } \mathscr{D}'((0,T)\times(L_2,L_3)) \tag{2.51}$$

to conclude

$$\begin{aligned} &m_2(x)(\partial_t w_n)^2-a_2(x)(\partial_x w_n)^2\\ &\quad\to m_2(x)(\partial_t w_\infty)^2-a_2(x)(\partial_x w_\infty)^2 \quad \text{in } \mathscr{D}'((\zeta/4,T-\zeta/4)\times(L_2,L_3)).\end{aligned} \tag{2.52}$$

By virtue of (1.7), (1.8), and (2.46), we find

$$\begin{aligned}&\int_{\zeta/2}^{T-\zeta/2}\left(\int_{L_2}^{L_2+\varepsilon_1}+\int_{L_3-\varepsilon_1}^{L_3}\right)\{(\partial_x w_n)^2+(\partial_t w_n)^2\}\,dx\,dt\le M\varepsilon_1\\ &\qquad\qquad\text{for all } n \text{ and all small } \varepsilon_1>0,\end{aligned} \tag{2.53}$$

$$\begin{aligned}&\int_{L_2+\varepsilon_1}^{L_3-\varepsilon_1}\left(\int_{\zeta/2}^{\zeta/2+\varepsilon_2}+\int_{T-\zeta/2-\varepsilon_2}^{T-\zeta/2}\right)\{(\partial_x w_n)^2+(\partial_t w_n)^2\}\,dt\,dx\leqq C(\varepsilon_1)\varepsilon_2\\ &\qquad\qquad\text{for all } n \text{ and all small } \varepsilon_2>0.\end{aligned} \tag{2.54}$$

Here, M and $C(\varepsilon_1)$ are positive constants.

Fix any small $\varepsilon>0$, and choose ε_1 so that

$$M\varepsilon_1<\varepsilon \qquad (M \text{ is the same as in (2.53)}) \tag{2.55}$$

and then, choose ε_2 so that

$$C(\varepsilon_1)\varepsilon_2<\varepsilon \qquad (C(\varepsilon_1) \text{ is the same as in (2.54)}). \tag{2.56}$$

There is a function $\Psi\in C_0^\infty((\zeta/2,T-\zeta/2)\times(L_2,L_3))$ such that $0\leqq\Psi\leqq1$ and $\Psi(t,x)=1$, for $(t,x)\in[\zeta/2+\varepsilon_2,T-\zeta/2-\varepsilon_2]\times[L_2+\varepsilon_1,L_3-\varepsilon_1]$.

By (2.52), it is clear that

$$\begin{aligned}&\lim_{n\to\infty}\int_{\zeta/2}^{T-\zeta/2}\int_{L_2}^{L_3}\Psi\{m_2(x)(\partial_t w_n)^2-a_2(x)(\partial_x w_n)^2\}\,dx\,dt\\ &\quad=\int_{\zeta/2}^{T-\zeta/2}\int_{L_2}^{L_3}\Psi\{m_2(x)(\partial_t w_\infty)^2-a_2(x)(\partial_x w_\infty)^2\}\,dx\,dt.\end{aligned} \tag{2.57}$$

Combining (2.53)-(2.57), we obtain

$$\begin{aligned}&\overline{\lim_{n\to\infty}}\left|\int_{\zeta/2}^{T-\zeta/2}\int_{L_2}^{L_3}\{m_2(x)(\partial_t w_n)^2-a_2(x)(\partial_x w_n)^2\}\,dx\,dt\right.\\ &\qquad\left.-\int_{\zeta/2}^{T-\zeta/2}\int_{L_2}^{L_3}\{m_2(x)(\partial_t w_\infty)^2-a_2(x)(\partial_x w_\infty)^2\}\,dx\,dt\right|\\ &\qquad<C\varepsilon \quad \text{where } C \text{ is a positive constant independent of } \varepsilon.\end{aligned} \tag{2.58}$$

Note that w_∞ also satisfies (2.53) and (2.54). Since ε is arbitrary, we have

$$\begin{aligned}&\lim_{n\to\infty}\int_{\zeta/2}^{T-\zeta/2}\int_{L_2}^{L_3}\{m_2(x)(\partial_t w_n)^2-a_2(x)(\partial_x w_n)^2\}\,dx\,dt\\ &\quad=\int_{\zeta/2}^{T-\zeta/2}\int_{L_2}^{L_3}\{m_2(x)(\partial_t w_\infty)^2-a_2(x)(\partial_x w_\infty)^2\}\,dx\,dt.\end{aligned} \tag{2.59}$$

In the meantime, it follows from (2.43) that

$$\lim_{n\to\infty}\int_0^T\int_{L_2}^{L_3}(\partial_t w_n)^2\,dx\,dt=0 \tag{2.60}$$

and thus

$$\int_{\zeta/4}^{T-\zeta/4}\int_{L_2}^{L_3}(\partial_t w_\infty)^2\,dx\,dt=0. \tag{2.61}$$

Now we see from (2.42), (2.59), (2.60), and (2.61) that

$$\int_{\zeta/2}^{T-\zeta/2}\int_{L_2}^{L_3}a_2(x)(\partial_x w_\infty)^2\,dx\,dt=1. \tag{2.62}$$

Since $\partial_t w_\infty\equiv 0$ in $(\zeta/4, T-\zeta/4)\times(L_2, L_3)$, it follows from Lemma 1.2 that

$$\lim_{x\to L_2+}a_2(x)\partial_x w_\infty(t,x)=\beta \quad \text{in } L^2(\zeta/2, T-\zeta/2), \tag{2.63}$$

for some constant β. By the same argument as for (2.30) and (2.31), we see that

$$\lim_{x\to L_2-}\partial_t w_\infty(t,x)=0 \quad \text{in } L^2(3\zeta/4, T-3\zeta/4), \tag{2.64}$$

$$\lim_{x\to L_2-}a_1(x)\,\partial_x w_\infty(t,x)=\beta \quad \text{in } L^2(3\zeta/4, T-3\zeta/4). \tag{2.65}$$

Let us consider the initial value problem:

$$-\begin{pmatrix}0, & 1\\ a_1(x)/m_1(x), & 0\end{pmatrix}\partial_x Y-\begin{pmatrix}0, & 0\\ \partial_x a_1(x)/m_1(x), & 0\end{pmatrix}Y=0, \tag{2.66}$$

$$Y(L_2)=\begin{pmatrix}\beta/a_1(L_2)\\ 0\end{pmatrix}. \tag{2.67}$$

Then, by Lemma 1.1, we conclude that

$$Y(x)=\begin{pmatrix}\partial_x w_\infty(t,x)\\ \partial_t w_\infty(t,x)\end{pmatrix} \quad \text{in } (\zeta, T-\zeta)\times(L_1, L_2). \tag{2.68}$$

Consequently, $\partial_t w_\infty\equiv 0$ in $(\zeta, T-\zeta)\times(L_1, L_3)$, and w_∞ satisfies, according to (2.50),

$$\langle w_\infty, \phi\rangle_2=0 \quad \text{for all } \phi\in H_0^1((L_1, L_3)), \tag{2.69}$$

which yields $w_\infty\equiv 0$ in $(\zeta, T-\zeta)\times(L_1, L_3)$. This contradicts (2.62), for $\partial_t w_\infty\equiv 0$ in $(\zeta/2, T-\zeta/2)\times(L_2, L_3)$. This completes the proof.

3. Final remarks. We have considered only a special combination of different bars. But it is obvious that the same method can be applied to any combination of several different bars at least one of which is damped.

In (2.3), we assumed that $b(x)\geqq\beta_3>0$, for all $x\in[L_2, L_3]$. If $b(x)\geqq\beta_3>0$, only for $x\in I\subset[L_2, L_3]$, then we can regard the bar as a combination of different bars: $[L_2, L_3]=I\cup$(one or two segments). Then, the same method can be still applied.

It is also easy to see that the boundary condition at one end can be replaced by the homogeneous Neumann boundary condition.

Note added in proof. After this paper was submitted, the author found the works of Ho [9] and Lagnese [10] which are closely related to our problem. They considered the case of smooth coefficients without discontinuity by different methods.

REFERENCES

[1] G. Chen, *Control and stabilization for the wave equation in a bounded domain*, SIAM J. Control Optim., 17 (1979), pp. 66-81.

[2] G. Chen, S. A. Fulling, F. J. Narcowich, and C. Qi, *An asymptotic average decay rate for the wave equation with variable coefficient viscous damping*, SIAM J. Appl. Math., 50 (1990), pp. 1341-1347.

[3] G. Chen, S. A. Fulling, F. J. Narcowich, and S. Sun, *Exponential decay of energy of evolution equations with locally distributed damping*, SIAM J. Appl. Math., 51 (1991), pp. 266-301.

[4] B. Dacorogna, *Weak Continuity and Weak Lower Semicontinuity of Nonlinear Functionals*, Lecture Notes in Mathematics, Vol. 922, Springer-Verlag, Berlin, New York, 1982.

[5] I. Lasiecka, J. L. Lions, and R. Triggiani, *Nonhomogeneous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl., 65 (1986), pp. 149-192.

[6] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, Berlin, New York, 1972.

[7] J. Rauch and M. Taylor, *Exponential decay of solutions to hyperbolic equations in bounded domains*, Indiana Univ. Math. J., 24 (1974), pp. 79-86.

[8] W. A. Strauss, *On continuity of functions with values in various Banach spaces*, Pacific J. Math., 19 (1966), pp. 543-551.

[9] L. F. Ho, *Exact controllability of the one-dimensional wave equation with locally distributed control*, SIAM J. Control Optim., 28 (1990), pp. 733-748.

[10] J. Lagnese, *Control of wave processes with distributed controls supported on a subregion*, SIAM J. Control Optim., 21 (1983), pp. 68-85.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 381-402, March 1991

ON A CONVEX PARAMETER SPACE METHOD FOR LINEAR CONTROL DESIGN OF UNCERTAIN SYSTEMS*

J. C. GEROMEL†, P. L. D. PERES†, AND J. BERNUSSOU‡

Abstract. This paper presents a new procedure for continuous and discrete-time linear control systems design. It consists of the definition of a *convex* programming problem in the parameter space that, when solved, provides the feedback gain. One of the most important features of the procedure is that additional design constraints are easily incorporated in the original formulation, yielding solutions to problems that have raised a great deal of interest within the last few years. This is precisely the case of the decentralized control problem and the quadratic stabilizability problem of uncertain systems with both dynamic and input uncertain matrices. In this last case, necessary and sufficient conditions for the existence of a linear stabilizing gain are provided and, to the authors' knowledge, this is one of the first numerical procedures able to handle and solve this interesting design problem for high-order, continuous-time or discrete-time linear models. The theory is illustrated by examples.

Key words. linear system control, linear programming, cutting-plane techniques, robust control

AMS(MOS) subject classification. 93C15

1. Introduction. The Linear-Quadratic Problem (LQP) has been extensively studied in the last decades. As a natural consequence of an enormous deal of work and interest, many properties of its solution came to light (see, for instance, [1] and the references therein). One of the most important properties states that the optimal solution of the LQP can be expressed as a linear feedback law which stabilizes the closed-loop system. It can be numerically determined by finding the positive-definite solution of a Riccati-type equation which also provides a quadratic Lyapunov function associated with the closed-loop system [1]. This important fact has served as a basis for the analysis of the dynamic behaviour of the closed-loop system when subjected to several classes of perturbations yielding many known robustness conditions [9], [21], [22].

However, almost all properties of the LQP, including those mentioned above, disappear when any additional design constraints are added. Among the problems raised, either the linearity of the optimal control law is lost or the associated problem to be solved in the parameter space (the elements of the matrix gain) does not present any property like convexity which obviously is an interesting one for optimization purposes [8], [15]. Furthermore, generally an initial feasible gain which stabilizes the system is asked for and this can be a difficult task depending on the strucure of the design constraints. In the case of the decentralized control problem, a tentative can be made from the solutions of the LQP associated to the isolated subsystems. Unfortunately, this procedure works only if the plant is weakly coupled [23] or if certain matching conditions are fulfilled [10].

More recently, the LQP has been used to determine stabilizing feedback gains for uncertain systems (see [24] and the references therein). When some matching conditions are fulfilled, the stabilizing gain is determined from a Riccati-type equation, provided

* Received by the editors July 5, 1989; accepted for publication (in revised form) June 1, 1990. This research has been supported in part by grant 301373/80 from Conselho Nacional de Desenvolvimento Científico e Tecnológico - CNPq/Brazil and CEFI/France.

† Laboratório de Análise Convexa—LAC/FEE UNICAMP, CP 6101, 13081, CAMPINAS, SP, Brazil.

‡ Laboratoire d'Automatique et d'Analyse des Systèmes du Centre National de la Recherche Scientific, Paris, France. 7, avenue du Colonel Roche, 31077, Toulouse Cedex, France et Greco Automatique, Centre National de la Recherche Scientific, Paris, France.

it admits a positive-definite solution [17]. However, it is important to keep in mind that in all cases we must handle a nonlinear equation that probably must be solved many times until it is decided whether or not there is a positive-definite solution [12], [17]. This can be a serious drawback mainly if high-order systems are under consideration.

In this paper we propose a convex parameter space method for linear system control design. It is based on the fact that the stabilizability property is related to the existence of a nonempty convex cone defined on the parameter space in the elements of a positive-definite matrix. Consequently, the feedback control synthesis derives from the global solution of a *convex* problem defined over the above-mentioned convex cone.[1]

The paper contributes in several directions. First, the convexity of the design problem opens the possibility of solving it, directly in the parameter space, by means of the most powerful tools available to date in the mathematical programming literature. Furthermore, additional convex design constraints are easily incorporated into the original problem. This fact allows us to solve the decentralized control problem for a wide class (called strongly decentralized) of linear dynamic systems. Necessary and sufficient conditions for its solvability are provided. Another contribution is related to the concept of quadratic stabilizability of uncertain systems [3]. Assuming the set of uncertainties is bounded and convex (an hyper-rectangle, for instance), the uncertain system is shown to be quadratically stabilizable if and only if the intersection of a finite number of convex cones is nonempty. This simplifies the results of [3] and enables us once again to find the Lyapunov function by means of a convex problem. The feedback gain is found directly from its solution which is numerically determined by constraints generation or by its dual the Dantzig–Wolfe decomposition principle (column generation) [13]. Finally, all the results above are generalized to discrete-time systems. In this case, the quadratic stabilizability problem is solved for the first time without assuming any special system structure or matching conditions.

The paper is organized as follows. The first sections are devoted to the analysis of continuous-time systems. In § 2 some definitions are introduced and the basic design problem is stated. Extensions for dealing with decentralized control and quadratic stabilization of uncertain systems are considered in § 3. Since one of the main control objectives is to get a small feedback gain, the index of the basic design problem is chosen for this purpose.[2] It will be defined in terms of an extended system model introduced in § 4. In the next section, several numerical aspects are considered. In particular, the iterative procedure for control design is stated and its convergence analysed.

In § 6 the same results for discrete-time systems are stated. Although there exist many different theoretical aspects when compared with the theory developed for continuous-time systems, it will be shown that the same numerical procedure can still be used. Finally in § 7 some examples are solved and simulation results are included and compared with others available in the literature. In § 8 we summarize the most important conclusions.

2. Definitions and problem statement. In this section we introduce the basic design problem written directly in the parameter space. Its relationship with the classical LQP

[1] For alternative approaches leading to convex programming problems, see [6] and [20]. Unfortunately, the parametrization introduced results in high-order compensators.

[2] In principle, other important performance functions can be considered, making use of the concept of majorization [19].

is investigated as well. Consider a continuous-time linear system

(1) (S): $\dot{x} = Ax(t) + Bu(t)$,

where $x(t) \in \Re^n$ is the state vector and $u(t) \in \Re^m$ is the control vector. The real matrices A and B are of appropriate dimensions. Let us define the following set of matrices:

$$\begin{aligned} \Sigma_0 &\triangleq \{-xx' \in \Re^{n\times n}: \|x\| = 1\}, \\ \Sigma_S &\triangleq \{2A'xx' \in \Re^{n\times n}: B'x = 0, \|x\| = 1\}, \end{aligned} \tag{2}$$

where $\|x\|$ denotes the Euclidean norm of $x \in \Re^n$. Denoting the inner product of matrices in the $\Re^{n\times n}$ as $\langle W, X\rangle \triangleq \text{Trace}\,(W'X)$, the set $\mathscr{C}$ will be important in further developments:

$$\mathscr{C} \triangleq \{W \in \Re^{n\times n}: \langle W, X\rangle < 0, X \in \Sigma\}, \tag{3}$$

where $\Sigma \triangleq \Sigma_0 \cup \Sigma_S$ and W is assumed to be real and symmetric . This set is strongly related to the stabilizability of system (S). This property can be easily verified by noting that:[3]

$$\langle W, X\rangle < 0: X \in \Sigma_0 \Leftrightarrow x'Wx > 0 \quad \forall x \in \Re^n \tag{4}$$

and a similar relation holds for $X \in \Sigma_S$, that is,

$$\langle W, X\rangle < 0: X \in \Sigma_S \Leftrightarrow x'(AW + WA')x < 0 \quad \forall x: B'x = 0. \tag{5}$$

Consequently, for $W \in \mathscr{C}$, both inequalities (4) and (5) hold simultaneously. In view of the above, the interpretation of $\mathscr{C}$ is quite simple. Indeed, (4) states that $W = W' > 0$ (positive definiteness) and (5) implies that all modes of (1) belonging to the nullspace of B' are asymptotically stable. Using a definition given in [25], we conclude that this is possible only if the pair (A, B) is stabilizable. More precisely, we have the following result.

THEOREM 1. *The set $\mathscr{C}$ defined in* (3) *is such that*

(a) *$\mathscr{C} \neq \emptyset$ if and only if (A, B) is stabilizable.*

(b) *$\mathscr{C}$ is a convex cone.*

(c) *If $W_* \notin \mathscr{C}$, there exists $X_* \in \Sigma$ such that the set*

$$\mathscr{C}_* \triangleq \{W \in \Re^{n\times n}: \langle W, X_*\rangle < 0\}$$

satisfies $W_* \notin \mathscr{C}_* \supset \mathscr{C}$.

Proof. The proof of part (a) is simple and is left to the reader. The proof that $\mathscr{C}$ is a convex set follows from its definition (3). Indeed, $\mathscr{C}$ is the intersection of an infinite number of open halfspaces defined by linear (and hence convex) inequalities. On the other hand, for any $W \in \mathscr{C}$, $\lambda W \in \mathscr{C}$, for all $\lambda > 0$, consequently $\mathscr{C}$ is a cone. To prove the last part, we define X_* as follows:

$$X_* \triangleq \arg\max \{\langle W_*, X\rangle: X \in \Sigma\}. \tag{6}$$

Since by assumption $W_* \notin \mathscr{C}$, then $\langle W_*, X\rangle \geqq 0$, which implies that $W_* \notin \mathscr{C}_*$. In addition, X_* being an element of Σ, it is obvious that $\mathscr{C}$ is a subset of $\mathscr{C}_*$. This proves the theorem.

Theorem 1 deserves some remarks. From the proof we verify that in fact $\mathscr{C}$ is an open convex set. However, being a cone, it can be rewritten as

$$\mathscr{C} = \{W \in \Re^{n\times n}: \langle W, X\rangle \leqq -\beta, X \in \Sigma\} \tag{7}$$

[3] For arbitrary matrices with appropriate dimension, Trace (AB) = Trace (BA).

with $\beta > 0$. Now, $\mathscr{C}$ is closed and no loss of generality is introduced by a particular choice of $\beta > 0$. Indeed, as will be seen in the sequel, the parameter β has no influence on the determination of a stabilizing feedback gain. However, it is important to note that in this case the set $\mathscr{C}_*$ must be changed into

$$\mathscr{C}_* = \{W \in \Re^{n\times n} : \langle W, X_* \rangle \leqq -\beta\}. \tag{8}$$

Based on the above discussion, we adopt throughout the text the normalized value $\beta = 1$.

Part (c) of Theorem 1 states that for each $W_* \notin \mathscr{C}$ there exists a separating hyperplane which separates W_* from $\mathscr{C}$. It is calculated from the optimal solution of problem (6). At first glance it appears difficult to solve. However, from the definition of Σ we readily see that

$$\max\{\langle W_*, X \rangle : X \in \Sigma\} = \max\{\rho_0, \rho_S\}, \tag{9}$$

where ρ_0 and ρ_S are the optimal values of the left-hand side of (9) with Σ replaced by Σ_0 and Σ_S, respectively. Using (2), we get

$$\rho_0 = -\lambda_m[W_*], \qquad \rho_S = \lambda_M[T'\Theta(W_*)T], \tag{10}$$

where $\Theta(W) \triangleq AW + WA'$, $T \in \Re^{n\times n-m}$ is an orthonormal matrix spanning the null-space of B' and $\lambda_m[\cdot]$, $\lambda_M[\cdot]$ denote the minimum and the maximum eigenvalues of $[\cdot]$, respectively. Obviously, the optimal solution $X_* \in \Sigma$ is determined from the normalized eigenvectors associated with ρ_0 or ρ_S.

Now, assume that the pair (A, B) is stabilizable and pick up from $\mathscr{C}$ a generic element W. From (4), we have $W = W' > 0$ and, using Finsler's lemma (see, for example, [17]) to (5), there exists a symmetric positive-definite matrix $R \in \Re^{m\times m}$ such that $\Theta(W) < BR^{-1}B'$ (see [4] for a particular choice of R). Consequently, the symmetric positive definite matrix $Q \in \Re^{n\times n}$

$$Q = W^{-1}[BR^{-1}B' - \Theta(W)]W^{-1} \tag{11}$$

is such that $P = W^{-1} > 0$ solves the following Riccati equation:

$$A'P + PA - PBR^{-1}B'P + Q = 0. \tag{12}$$

This shows that the linear time-invariant feedback gain $K = R^{-1}B'W^{-1}$ is such that the control law $u(t) = -Kx(t)$ stabilizes (1) asymptotically and is optimal with respect to the quadratic index

$$\min_u \int_0^\infty \{x(t)'Qx(t) + u(t)'Ru(t)\}\,dt \tag{13}$$

with $V(x) = x'Px$ being a Lyapunov function for the closed-loop system. From the results above, we conclude that there is a one-to-one correspondence between each element of $\mathscr{C}$ and a pair of positive-definite matrices (Q, R) which defines a standard LQP. That is, the condition $\mathscr{C} \neq \varnothing$ and the existence of a positive-definite solution for (12) are equivalent.

However, from a numerical point of view, the above equivalence no longer takes place. In the classical LQ design, matrices Q and R are chosen and (12) is solved for $P > 0$, giving rise to the optimal feedback gain. As discussed in the Introduction, the nonlinear nature of (12) imposes serious difficulties for the inclusion in the LQP of additional design constraints.

Theorem 1 allows us to work directly in the parameter space $\Re^{n\times n}$. A particular stabilizing gain (and optimal in a certain sense) can be determined from

$$\min\{f(W) : W \in \mathscr{C}\}, \tag{14}$$

where $f(\cdot)$: $\Re^{n\times n}\to\Re$ is a convex matrix-valued function. Again, in view of Theorem 1, (14) is a feasible *convex* problem whenever the pair (A, B) is stabilizable. As will be seen in the sequel, all machinery available to date in the mathematical programming literature can be used to solve it. Furthermore, additional design constraints like decentralization or parameter uncertainties are relatively easily incorporated in the basic structure.

Keeping this fact in mind, the final choice of the index $f(W)$ follows from a simple design strategy: Define convex constraints which reflect the designer objectives (a degree of stability $\alpha>0$ may also be included by simple substitution of A by $A+\alpha\mathbf{I}$ in Σ_S) and choose $f(W)$ such as the feedback gain determined from the optimal solution of (14) be "small." This procedure will be deeply analysed in § 4.

3. Decentralized and uncertain systems. First, we analyse the decentralized control problem [5], [24]. To this end, we assume that in (1), B is a block-diagonal matrix, that is,

$$B=B_D=\text{block-diagonal}\{B_1,\cdots,B_N\},\tag{15}$$

where $B_i\in\Re^{n_i\times m_i}$, $i=1\cdots N$. For convenience we introduce the following notation. A subscript "D" in a matrix indicates that it is block-diagonal, each block with appropriate dimensions. The set $\mathscr{C}$ is now replaced by

$$\mathscr{C}_D\triangleq\{W_D\in\Re^{n\times n}:\langle W_D,X\rangle<0,X\in\Sigma\}\tag{16}$$

and we define a class of decentralized dynamic systems.

DEFINITION 1. The system (S) is called *strongly decentralized* if there exist $W_D=W'_D>0$ and K_D such that $(A-BK_D)'W_D^{-1}+W_D^{-1}(A-BK_D)<0$.

Note that for this class not only the system must be stabilized by a decentralized gain but, in addition, the Lyapunov equation associated with the closed-loop matrix must have a block-diagonal solution. It is important to note that almost all the results available in the decentralized control literature deal with this class of systems.

THEOREM 2. *The system* (S) *is strongly decentralized if and only if* $\mathscr{C}_D\neq\emptyset$.

Proof. The "only if" part is proved assuming that there exist W_D and K_D satisfying Definition 1. In this case $(A-BK_D)W_D+W_D(A-BK_D)'<0$. Multiplying this inequality by x' on the left and x on the right, for $x\in\Re^n$ arbitrary, we get $\langle W_D,X\rangle<0$: $X\in\Sigma_S$. Since W_D is positive definite and block-diagonal, we conclude that $W_D\in\mathscr{C}_D$.

The "if" part is proved by construction. Take any $W_D\in\mathscr{C}_D$ and determine R_D such that $\Theta(W_D)<BR_D^{-1}B'$ (this is always possible by Finsler's lemma). Define $K_D=R_D^{-1}B'_DW_D^{-1}$ and observe that

$$\begin{aligned}(A-BK_D)'W_D^{-1}&+W_D^{-1}(A-BK_D)\\&=A'W_D^{-1}+W_D^{-1}A-2W_D^{-1}BR_D^{-1}B'W_D^{-1}\\&\leqq W_D^{-1}\{\Theta(W_D)-BR_D^{-1}B'\}W_D^{-1}<0.\end{aligned}\tag{17}$$

Note that by assumption $B=B_D$, and the theorem follows from Definition 1.

As a final comment, we note that Theorem 1, as well as all remarks after it, remains valid if we replace the matrices by their decentralized representation. Obviously, the difference between both theorems stems from the decentralized nature of the matrices involved. In fact, $\mathscr{C}_D$ could be defined from $\mathscr{C}$ by adding a number of linear constraints which reduce to zero all off-diagonal elements of $W\in\Re^{n\times n}$. As a conclusion, we note that the optimal solution of (14) for decentralized control synthesis is easier to calculate than the former one since fewer variables are involved.

Now we turn our attention to uncertain systems control. First, we consider the case in which only the A matrix is uncertain. The general case will be analysed in the next section. Suppose the linear system (1) is such that $A \in D_A$, where

$$D_A \triangleq \left\{ A \in \Re^{n \times n} \colon A = \sum_{i=1}^{N} \lambda_i A_i,\ \lambda_i \geqq 0,\ \sum_{i=1}^{N} \lambda_i = 1 \right\}. \tag{18}$$

Obviously, D_A is a convex and bounded domain. It is defined by a convex combination of the "extreme" matrices A_i, $i = 1 \cdots N$. We claim that this representation of D_A is quite general. A particular and important case corresponds to linear systems where the coefficients are known up to a certain precision defined by bounding them from below and above (interval matrices).

DEFINITION 2. A collection of linear systems defined by $(A_1, \cdots, A_N, B)$ is quadratically stabilizable via linear feedback control if there exist $W = W' > 0$ and K such that $(A_i - BK)'W^{-1} + W^{-1}(A_i - BK) < 0$ for all $i = 1 \cdots N$.

Defining the convex cone (see equation (3)):

$$\mathscr{C}_U \triangleq \{ W \in \Re^{n \times n} \colon \langle W, X \rangle < 0, X \in \Sigma_U \}, \tag{19}$$

where $\Sigma_U \triangleq \Sigma_0 \cup \Sigma_1 \cup \cdots \cup \Sigma_N$. The sets Σ_i, $i = 1, \cdots N$ are given by (2) with A replaced by A_i, $i = 1 \cdots N$. We have the following result.

THEOREM 3. *The collection* $(A_i, \cdots, A_N, B)$ *is quadratically stabilizable via linear control if and only if* $\mathscr{C}_U \neq \varnothing$.

Proof. The "only if" part is obvious and thus omitted (see the proof of Theorem 2). For the "if" part take an arbitrary $W \in \mathscr{C}_U$, define $\Theta_i(W) \triangleq A_i W + WA_i'$, and determine positive definite matrices R_i such that $\Theta_i(W) < BR_i^{-1}B'$, $i = 1 \cdots N$.

It is clear that any matrix $R = R' > 0$ such that $R^{-1} \geqq R_i^{-1}$ for all $i = 1 \cdots N$ satisfies $\Theta_i(W) < BR^{-1}B'$. Setting $K = R^{-1}B'W^{-1}$, we get

$$\begin{aligned} &(A_i - BK)'W^{-1} + W^{-1}(A_i - BK) \\ &= A_i'W^{-1} + W^{-1}A_i - 2W^{-1}BR^{-1}B'W^{-1} \\ &\leqq W^{-1}\{\Theta_i(W) - BR^{-1}B'\}W^{-1} < 0, \qquad i = 1 \cdots N \end{aligned} \tag{20}$$

and the theorem is proved by Definition 2.

It is interesting to observe that with the positive-definite matrices

$$Q_i = W^{-1}[BR^{-1}B' - \Theta_i(W)]W^{-1}, \qquad i = 1 \cdots N \tag{21}$$

all Riccati equations

$$A_i'P_i + P_iA_i - P_iBR^{-1}B'P_i + Q_i = 0 \tag{22}$$

have the same positive-definite solution, namely, $P_i = W^{-1}$, $i = 1 \cdots N$. Consequently, to each pair (A_i, B) it is possible to define a LQP which provides a linear feedback gain independent of the index $i = 1 \cdots N$.

On the other hand, we can verify that the set Σ_U does not introduce any fundamental additional difficulty for the determination of the separating hyperplane (see Theorem 1). Suppose that $W_* \notin \mathscr{C}_U$; then

$$\max \{ \langle W_*, X \rangle \colon X \in \Sigma_U \} = \max \{ \rho_0, \rho_1, \cdots, \rho_N \}, \tag{23}$$

where ρ_0 is given in (10) and

$$\rho_i = \lambda_M[T'\Theta_i(W_*)T], \qquad i = 1 \cdots N. \tag{24}$$

The optimal solution $X_* \in \Sigma_U$ is determined from the normalized eigenvectors associated to ρ_0 and $\rho_1 \cdots \rho_N$.

THEOREM 4. *Assume $A \in D_A$. The system* (S) *is quadratically stabilizable via linear control if and only if the same holds for the collection* $(A_i, \cdots, A_N, B)$.

Proof. The necessity is obvious since $A_i \in D_A$. The sufficiency is proved by taking $K = R^{-1}B'W^{-1}$ where $W \in \mathscr{C}_U$ and R is the matrix defined in the proof of Theorem 3. For all $A \in D_A$, we have (remember that $\lambda_i \geqq 0$ and $\sum_{i=1}^N \lambda_i = 1$)

$$\begin{aligned}(25)\qquad &(A-BK)'W^{-1}+W^{-1}(A-BK)\\ &=\left(\sum_{i=1}^N \lambda_i A_i - BK\right)' W^{-1} + W^{-1}\left(\sum_{i=1}^N \lambda_i A_i - BK\right)\\ &=\sum_{i=1}^N \lambda_i\{(A_i-BK)'W^{-1}+W^{-1}(A_i-BK)\}\\ &\leqq -\sum_{i=1}^N \lambda_i Q_i < 0,\end{aligned}$$

where the last inequality follows from (20) and (21). This proves the theorem proposed.

In view of both Theorems 3 and 4, we conclude that the uncertain system (S) is quadratically stabilizable via linear control if and only if all extreme matrices which define the uncertain domain D_A are quadratically stabilizable via linear control. This improves the results of [16] (mainly because it can be generalized to cope with uncertainties in the input matrix too) since only a finite (N) number of matrices have to be handled.

We want to stress that $\mathscr{C}_U$ is a convex cone for which separating hyperplanes are easy to determine (23). Based on this property, we propose in § 5 a simple and efficient algorithm which verifies whether or not $\mathscr{C}_U \neq \emptyset$ and an affirmative answer provides the stabilizing feedback again.

4. The extended system. In this section we introduce the extended system model associated to system (S). The idea was first proposed in [2] and, as is seen below, it allows us to solve the problems stated previously in a much more general framework. For convenience, through this section we note $p = n + m$.

The extended system associated to (S) is defined by

$$(26)\qquad (\mathrm{E}):\quad \dot z = Fz(t) + Gv(t),$$

where $z(t) \in \mathfrak{R}^p$ is the extended state vector and $v(t) \in \mathfrak{R}^m$ is the control vector. Matrices $F \in \mathfrak{R}^{p\times p}$ and $G \in \mathfrak{R}^{p\times m}$ are given by

$$(27)\qquad F = \begin{bmatrix} A & -B \\ 0 & 0 \end{bmatrix}, \qquad G = \begin{bmatrix} 0 \\ I \end{bmatrix}.$$

As in § 2 we define the sets of matrices

$$(28)\qquad \begin{aligned}\Sigma_0 &\triangleq \{-zz' \in \mathfrak{R}^{p\times p}:\ G'z = 0,\ \|z\| = 1\},\\ \Sigma_E &\triangleq \{2F'zz' \in \mathfrak{R}^{p\times p}:\ G'z = 0,\ \|z\| = 1\}.\end{aligned}$$

Note that Σ_0 in (28) is slightly different from the set defined in (2). As a consequence, the elements of the cone

$$(29)\qquad \mathscr{C} \triangleq \{\mathscr{W} \in \mathfrak{R}^{p\times p}:\ \langle \mathscr{W}, Z\rangle < 0,\ Z \in \Sigma\}$$

with $\Sigma \triangleq \Sigma_0 \cup \Sigma_E$ are not necessarily positive-definite matrices.

COROLLARY 5. *The set $\mathscr{C}$ defined in* (29) *has the properties stated in Theorem* 1.

Proof. Only part (a) is proved. For the proof of parts (b) and (c) see Theorem 1. Assuming (A, B) is stabilizable, there exist matrices $W = W' > 0$ and K with appropriate dimensions such that

$$(A - BK)'W^{-1} + W^{-1}(A - BK) < 0. \tag{30}$$

Forming the matrix $\mathscr{W} \in \Re^{p \times p}$ as

$$\mathscr{W} = \begin{bmatrix} W & W'K' \\ KW & ? \end{bmatrix} \tag{31}$$

and, using the definition of $\mathscr{C}$, we conclude that $\mathscr{W} \in \mathscr{C}$. Conversely, assume that $\mathscr{W} \in \mathscr{C}$. Partitioning it as

$$\mathscr{W} = \begin{bmatrix} W_1 & W_2 \\ W_2' & W_3 \end{bmatrix}, \tag{32}$$

we immediately note that with $W = W_1$ and $K = W_2' W_1^{-1}$, (30) holds. This proves the first part of the corollary.

From the above we conclude that no condition is required for the matrix $W_3 \in \Re^{m \times m}$ and, more importantly, for any $\mathscr{W} \in \mathscr{C}$ it is always possible to choose W_3 such that $\mathscr{W} \in \mathscr{C}$ and $\mathscr{W} = \mathscr{W}' \geqq 0$.

Consequently, there is no loss of generality if we redefine $\mathscr{C}$ as being (see the discussion concerning the closure of $\mathscr{C}$ just after Theorem 1)

$$\mathscr{C} \triangleq \{\mathscr{W} = \mathscr{W}' \geqq 0 \colon \langle \mathscr{W}, Z \rangle \leqq -1, Z \in \Sigma\}. \tag{33}$$

Now, assume that $\mathscr{W} \in \mathscr{C}$ is partitioned according to (32). Two facts are of great importance. First, $W_1 \geqq \mathbf{I}$ and, second, the positive definiteness of $\mathscr{W}$ implies that $W_3 \geqq W_2' W_1^{-1} W_2$. For all $\mathscr{W} \in \mathscr{C}$ the corresponding stabilizing gains satisfy

$$KK' \leqq KW_1K' = W_2' W_1^{-1} W_2 \leqq W_3 = G'\mathscr{W}G, \tag{34}$$

which implies that $\|K\|^2 \leqq \lambda_M[G'\mathscr{W}G]$. We can select a small stabilizing gain from the solution of the basic design problem (14) with

$$f(\mathscr{W}) \triangleq \lambda_M[G'\mathscr{W}G]. \tag{35}$$

It is important to note that $f(\mathscr{W})$ is a convex (but not everywhere differentiable) function. Indeed, its epigraph [13] can be written as

$$\operatorname{epi} f \triangleq \{(\rho, \mathscr{W}) \colon \rho \geqq \langle \mathscr{W}, Z \rangle, Z \in \Sigma_0^{\perp}\}, \tag{36}$$

where $\Sigma_0^{\perp}$ is orthogonal to Σ_0 defined in (28).[4] At this point, it is obvious that the basic design problem can be rewritten as

$$\min \{\rho \colon (\rho, \mathscr{W}) \in \operatorname{epi} f, \mathscr{W} \in \mathscr{C}\}. \tag{37}$$

In the next section we propose an algorithm based on the dual-simplex procedure able to detect whether $\mathscr{C} \neq \varnothing$ and in the affirmative determine a stabilizing feedback gain with small norm (in the sense discussed above).

Some remarks are now in order. A decentralized optimal gain can be obtained from (37) by simply imposing on $\mathscr{W}$ the structure

$$\mathscr{W} = \begin{bmatrix} W_{1D} & W_{2D} \\ W_{2D}' & ? \end{bmatrix}. \tag{38}$$

[4] $\Sigma_0^{\perp}$ is given by $\Sigma_0^{\perp} \triangleq \{zz' \in \Re^{p \times p} \colon z \in \text{Range}(G), \|z\| = 1\}$. Note that for all $X \in \Sigma_0$ and $Y \in \Sigma_0^{\perp} \Rightarrow \langle X, Y \rangle = 0$.

In this case the decentralized control problem can be solved in a more general framework since we do not need to assume a special structure for the input matrix, that is, $B = B_D$.

Another important remark concerns the function $f(\mathcal{W})$ defined in (35). Indeed, problem (37) can be significantly simplified by imposing on $\mathcal{W}$ the particular structure

$$\mathcal{W} = \begin{bmatrix} W_1 & W_2 \\ W_2' & \rho \mathbf{I} \end{bmatrix}. \tag{39}$$

In this case, $f(\mathcal{W}) = \rho$, and (37) simplifies to

$$\min \{\rho \colon \mathcal{W} \in \mathcal{C}\}. \tag{40}$$

We are now in position to generalize our previous results to uncertain systems with both A and B uncertain matrices. To this end, we assume that A and B belong to polyhedral convex domain with $n_E(A)$ and $n_E(B)$ extreme matrices, respectively. In the same way, we assume that the extended system (E) is such that $F \in D_F$ where

$$D_F \triangleq \left\{ F \in \Re^{p \times p} \colon F = \sum_{i=1}^{M} \lambda_i F_i, \lambda_i \geqq 0, \sum_{i=1}^{M} \lambda_i = 1 \right\} \tag{41}$$

and $M = n_E(A) \times n_E(B)$. Following Definition 2, we say that the collection of linear systems $\{(A_i, B_i), i = 1 \cdots M\}$ is quadratically stabilizable via linear feedback control if there exist $W = W' > 0$ and K such that $(A_i - B_i K)' W^{-1} + W^{-1}(A_i - B_i K) < 0$ for all $i = 1 \cdots M$. As before we define the convex cone

$$\mathcal{C}_U \triangleq \{\mathcal{W} \in \Re^{p \times p} \colon \langle \mathcal{W}, Z \rangle < 0, Z \in \Sigma_U\}, \tag{42}$$

where $\Sigma_U \triangleq \Sigma_0 \cup \Sigma_1 \cup \cdots \cup \Sigma_M$. The sets Σ_i $i = 1 \cdots M$ are given by (28) with F replaced by F_i $i = 1 \cdots M$.

COROLLARY 6. *The collection* $\{(A_i, B_i), i = 1 \cdots M\}$ *is quadratically stabilizable via linear control if and only if* $\mathcal{C}_U \neq \emptyset$.

Proof. The "if" part closely follows the proof of Corollary 5. For the proof of the "only if" part, take any $\mathcal{W} \in \mathcal{C}_U$. Partitioning it as (32), we have

$$A_i W_1 + W_1 A_i' < W_2 B_i' + B_i W_2', \qquad i = 1 \cdots M. \tag{43}$$

On the other hand, with $W = W_1$ and $K = W_2' W_1^{-1}$, simple calculations yield

$$(A_i - B_i K)' W^{-1} + W^{-1}(A_i - B_i K) = W^{-1}[A_i W_1 + W_1 A_i' - W_2 B_i' - B_i W_2'] W^{-1} < 0. \tag{44}$$

Since the above inequality holds for all $i = 1 \cdots M$, the corollary is proved.

COROLLARY 7. *Assume that* $F \in D_F$. *The system* (S) *is quadratically stabilizable via linear control if and only if the same holds for the collection* $\{(A_i, B_i), i = 1 \cdots M\}$.

Proof. Once again the necessity is obvious. For the sufficiency, suppose the collection $\{(A_i, B_i), i = 1 \cdots M\}$ satisfies Corollary 6. In this case $\mathcal{C}_U \neq \emptyset$. After partitioning $\mathcal{W} \in \mathcal{C}_U$, setting $W = W_1$, $K = W_2' W_1^{-1}$, and noting that for each $F \in D_F$ there exist $\lambda_i \geqq 0$, $\sum_{i=1}^{M} \lambda_i = 1$ such that

$$(A, B) = \sum_{i=1}^{M} \lambda_i (A_i, B_i), \tag{45}$$

we immediately obtain for all $F \in D_F$

$$(A - BK)' W^{-1} + W^{-1}(A - BK) = \sum_{i=1}^{M} \lambda_i \{(A_i - B_i K)' W^{-1} + W^{-1}(A_i - B_i K)\} < 0, \tag{46}$$

where the last inequality follows from (44). This concludes the proof of the corollary.

We close this section with some remarks. From the definition of the extended system (E), it has been possible to determine an objective function for the basic design problem in such a way that a small feedback gain is provided by its optimal solution. This is an important practical aspect mainly because many other design requirements can be imposed by a proper choice of additional convex constraints.

The analysis of uncertain systems stability has been possible in the general case corresponding to A and B uncertain matrices. The results reported here generalize those available in the literature in several directions. First, in contrast with what is done in [16], we provided necessary and sufficient conditions for stabilizability of a collection of multivariable systems by means of linear feedback. Numerically speaking, the result of Corollary 7 is important. Indeed, the "abstract condition" given in [3] can be effectively solved with an important reduction of the computational burden since only a finite number of "extreme" models have to be considered.

5. Numerical procedure. In this section we present an algorithm for solving the convex programming problem[5]

$$\min\{f(\mathcal{W}): \mathcal{W}\in\mathcal{C}\}, \tag{47}$$

where $f(\cdot)$ is convex and $\mathcal{C}$ is a convex cone. From the results presented before, a particular design problem can be considered by a particular choice of the cone $\mathcal{C}$ and the structure of $\mathcal{W}\in\Re^{p\times p}$. The algorithm must be able to detect whether $\mathcal{C}\neq\emptyset$ and in the affirmative case provide $\mathcal{W}^*\in\mathcal{C}$ such that $f(\mathcal{W}^*)\leqq f(\mathcal{W})$, for all $\mathcal{W}\in\mathcal{C}$.

In our context, the more general problem to be solved can be stated as

$$\begin{aligned}&\min \rho\\ &\langle\mathcal{W}, Z\rangle\leqq\rho \quad \text{for all } Z\in\Sigma_0^{\perp},\\ &\langle\mathcal{W}, Z\rangle\leqq-\beta_i \quad \text{for all } Z\in\Sigma_i, \quad i=0\cdots M+1,\end{aligned} \tag{48}$$

where $\beta_i=1$, $i=0\cdots M$. The constraint corresponding to $i=M+1$ imposes $\mathcal{W}\geqq 0$. Consequently, we set $\beta_{M+1}=0$ and

$$\Sigma_{M+1}\triangleq\{-zz'\in\Re^{p\times p}: \|z\|=1\}. \tag{49}$$

For convenience, matrices $\mathcal{W}\in\Re^{p\times p}$ and $\Theta_i(\mathcal{W})=F_i\mathcal{W}+\mathcal{W}F_i'\in\Re^{p\times p}$ are partitioned according to (32), that is,

$$\mathcal{W}=\begin{bmatrix}W_1 & W_2\\ W_2' & W_3\end{bmatrix}, \quad \Theta_i(\mathcal{W})=\begin{bmatrix}\Theta_{1i}(\mathcal{W}) & ?\\ ? & ?\end{bmatrix}, \quad i=1\cdots M, \tag{50}$$

where $W_1\in\Re^{n\times n}$ and $\Theta_{1i}(\mathcal{W})\in\Re^{n\times n}$. Note that for each $\mathcal{W}\notin\mathcal{C}$ a separating hyperplane between $\mathcal{W}$ and $\mathcal{C}$ is readily calculated (see Theorem 1).

- *Step* 1. Set the iteration index $l=0$ and

$$\mathcal{W}^l=\begin{bmatrix}\mathbf{I} & 0\\ 0 & 0\end{bmatrix}.$$

- *Step* 2. Define the constants ρ_i, $i=0,\cdots,M+1$ as being

$$\begin{aligned}\rho_0&=-\lambda_m[W_1^l],\\ \rho_i&=\lambda_M[\Theta_{1i}(\mathcal{W}^l)], \qquad i=1\cdots M,\\ \rho_{M+1}&=-\lambda_m[\mathcal{W}^l]\end{aligned}$$

[5] See also [18]. The difficulty is that $\mathcal{C}$ is not known explicitly.

and determine $\omega^* = \max\{\rho_i + \beta_i : i = 0, \cdots, M+1\}$. If $\omega^* \leqq 0$ go to Step 6. Otherwise set $i^* = i$ such that $\rho_i + \beta_i = \omega^*$ and go to next step.

- *Step* 3. Set $\gamma^l = -\beta_{i^*}$ and determine $Z^l \in \Sigma_{i^*}$ from the normalized eigenvector associated to ρ_{i^*}. Clearly, z^l has the form $z^l = [(x^l)' \, 0]'$ if $0 \leqq i^* \leqq M$.
- *Step* 4. Determine the normalized eigenvector $y^l \in \Re^m$ associated to $\lambda_M[W_3^l]$. Set $z^l = [0 \, (y^l)']'$, $Y^l = (z^l)(z^l)'$ and solve by any applicable method the linear programming problem

$$\begin{aligned} &\min \rho \\ &\langle \mathcal{W}, Y^j \rangle \leqq \rho, \\ &\langle \mathcal{W}, Z^j \rangle \leqq \gamma^j, \end{aligned} \qquad j = 0, \cdots, l. \tag{51}$$

- *Step* 5. If (51) is unfeasible then $\mathcal{C} = \varnothing$, **stop**. Otherwise, let $\mathcal{W}^{l+1}$ be its optimal solution. Make $l \leftarrow l+1$ and go back to Step 2.
- *Step* 6. $\mathcal{W}^l \in \mathcal{C}$. Determine the feedback gain $K = (W_2^l)'(W_1^l)^{-1}$.

The algorithm above works with the relaxed version of problem (48). Thanks to its convexity, at each iteration two linear constraints are added to the relaxed master problem (51) which approximates better both the epi f and the cone $\mathcal{C}$. Each linear constraint is defined in terms of Z^l, which satisfies

$$Z^l = \arg\max\{\langle \mathcal{W}^l, Z\rangle : Z \in \Sigma\}. \tag{52}$$

This fact contributes to reducing the number of iterations since the rate of convergence of this class of algorithms depends mainly on the deepness of the cut associated to each separating hyperplane.

On the other hand, calling ρ^{l+1} the minimum value of the index in (51), the dual representation of its first l constraints gives for some $\mu_j \geqq 0$, $\sum_{j=1}^{l} \mu_j = 1$:

$$\begin{aligned} \rho^{l+1} &= \sum_{j=0}^{l} \mu_j \langle \mathcal{W}^{l+1}, Y^j \rangle \\ &= \left\langle \mathcal{W}^{l+1}, \sum_{j=0}^{l} \mu_j Y^j \right\rangle. \end{aligned} \tag{53}$$

Consequently, the relaxed master problem approximates $f(\mathcal{W}) \cong \langle \mathcal{W}, Y\rangle$ where Y is iteratively determined. Note that in the previous version of the algorithm, it stops when some feasible $\mathcal{W} \in \mathcal{C}$ is reached.

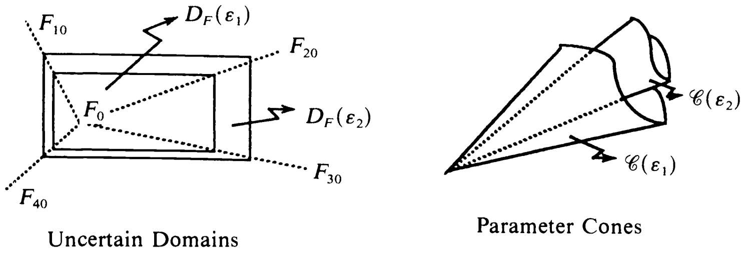
It is obvious that both structures (38) and (39) can be imposed to solve problems with decentralized constraints and simpler objective function. In the last case, (51) simplifies considerably to

$$\min\{\rho : \langle \mathcal{W}, Z^j \rangle \leqq \gamma^j, j = 1, \cdots, l\}. \tag{54}$$

The algorithm is sufficiently general to cope with other important problems. Indeed, the generalization to handle D_F convex but not necessarily polyhedral is immediate. The same occurs to the problem of synthesizing a linear feedback control so as to maximize the uncertainty parameter region [11]. Consider $D_F(\varepsilon)$, $\varepsilon \geqq 0$, the uncertainty domain defined in (41) for $F_i = F_0 + \varepsilon F_{1i}$, $i = 1 \cdots M$ (see Fig. 1) and assume that $\mathcal{C}(\varepsilon) \neq \varnothing$ for $\varepsilon = 0$.

In this case, the object function of (48) depends on $\varepsilon \in \Re$ and for each $\varepsilon \geqq 0$ it can be calculated by our procedure. The problem is to find ε for which $\mathcal{C}(\varepsilon) = \varnothing$ or $\rho(\varepsilon) = +\infty$. Since $\rho(\varepsilon)$ measures the magnitude of the feedback gain, from a practical point of view we have to solve

$$\max\{\varepsilon : \rho(\varepsilon) \leqq \rho_{\max}\}, \tag{55}$$



Uncertain Domains

Parameter Cones

FIG. 1. *Geometric interpretation of* $D_F(\varepsilon)$ *and* $\mathscr{C}(\varepsilon)$ *for* $\varepsilon_2 > \varepsilon_1$.

which needs in addition only a one-dimensional search. We claim that this approach appears to be more efficient than the one proposed in [11] since it avoids the calculation of the control gain itself which is determined only at the end of the iterative process.

THEOREM 8. *Suppose the algorithm generates the sequence of matrices* $\mathscr{W}^l$, $l = 0, 1, \cdots$. *Then, the following hold*:

(a) *If* $\mathscr{C} \neq \emptyset$, *any limit matrix of this sequence solves* (47).

(b) *If* $\mathscr{C} = \emptyset$, *an iteration index l will exist for which the linear problem* (51) *at Step* 4 *is unfeasible.*[6]

Proof. The proof is based on the convexity of problem (47). It follows the same pattern as that presented in [7] (see also [14]).

With regard to the convergence, Step 4 is of great importance. Since only two (or even one if (54) is under consideration) constraints are added to the linear programming problem (51), one of the best ways to solve it is using the DUAL-SIMPLEX procedure. Indeed, at any further iteration, the basis matrix is readily determined from the one of the previous iteration [13].

On the other hand, the global convergence stated in Theorem 8 is not destroyed if the algorithm is arranged by discarding the nonbinding constraints at the end of each iteration. Using the fact that $\mathscr{W}$ is symmetric, at most $p(p+1)/2$ linearly indepedent constraints have to be retained [14].

Of course this is not essential since the sequence ρ^l, $l = 0, 1, \cdots$ will still be monotonically nondecreasing (and bounded in the case $\mathscr{C} \neq \emptyset$) but certainly contributes for the global efficiency of the algorithm proposed.

6. Discrete-time systems design. We now focus our attention on discrete-time system given by

$$(56) \qquad (\mathrm{S}): \quad x(t+1) = Ax(t) + Bu(t),$$

where $x(t) \in \mathfrak{R}^n$ and $u(t) \in \mathfrak{R}^m$ are the state and control vectors. An important observation is that due to some nonlinear relationships introduced by the discrete-time Lyapunov equation, the feedback control design cannot be accomplished in the original space ($\mathfrak{R}^n$). For this reason, we immediately define the extended system associated to (S):

$$(57) \qquad (\mathrm{E}): \quad z(t+1) = Fz(t) + Gv(t),$$

where $z(t) \in \mathfrak{R}^p$, $v(t) \in \mathfrak{R}^m$ and F and G are given by (27). For convenience, throughout this section we use the same notation introduced in § 4.

[6] In practice, problem (51) is also declared unfeasible if, for some index l, $\rho^l > \rho_{\max}$, where $\rho_{\max}$ is a sufficiently large positive parameter.

The set Σ_E defined in (28) is now redefined as

$$\Sigma_E \triangleq \{F'zz'F - zz' \in \Re^{p\times p}\colon G'z = 0, \|z\| = 1\} \tag{58}$$

while the cone $\mathscr{C}$ defined in (29) moves to

$$\mathscr{C} \triangleq \{\mathscr{W} = \mathscr{W}' \geqq 0 \in \Re^{p\times p}\colon \langle \mathscr{W}, Z\rangle < 0, Z \in \Sigma\}, \tag{59}$$

where, as before, $\Sigma \triangleq \Sigma_0 \cup \Sigma_E$. From (59) we note the main difference between discrete-time and continuous-time linear system control design. Indeed, in the continuous-time case, the nonnegativeness constraint $\mathscr{W} \geqq 0$ has been introduced without loss of generality. On the contrary, in the discrete-time case, it is crucial to get the next results.

THEOREM 9. *The set $\mathscr{C}$ defined in* (59) *has the properties stated in Theorem* 1.

Proof. Once again only part (a) will be proved. The remaining ones are simple (for details see the proof of Theorem 1). Suppose first that (A, B) is stabilizable; then there exist matrices $W = W' > 0$ and K with appropriated dimensions such as

$$(A - BK)W(A - BK)' - W < 0. \tag{60}$$

Defining the matrix $\mathscr{W} \in \Re^{p\times p}$ as

$$\mathscr{W} = \begin{bmatrix} W & WK' \\ KW & KWK' \end{bmatrix} \tag{61}$$

and using the definition of $\mathscr{C}$, we conclude that indeed, $\mathscr{W} \in \mathscr{C}$. Conversely, assume that $\mathscr{W} \in \mathscr{C}$ is partitioned according to (32) and define $W = W_1$ and $K = W_2'W_1^{-1}$. Since any $z \in \Re^p$ such that $G'z = 0$ can be written as $z' = [x' \ \ 0]$ with $x \in \Re^n$ we have

$$x'[(A - BK)W(A - BK)' - W]x = z'[F\mathscr{W}F' - \mathscr{W}]z - x'B(W_3 - W_2'W_1^{-1}W_2)B'x. \tag{62}$$

The first term on the right-hand side of (62) is strictly negative because $\mathscr{W} \in \mathscr{C}$ and the second one is nonnegative because all feasible $\mathscr{W}$ are nonnegative definite matrices. Consequently, (60) holds and the theorem is proved.

For optimization purposes, the theorem above states that $\mathscr{C}$ is a convex cone which can be replaced by (33). The feedback gain is bounded below by $f(\mathscr{W})$ defined in (35) in which case the basic design problem (37) can be still used. Furthermore, a decentralized gain can be calculated (if any) by simply setting to $\mathscr{W}$ the structure (38).

Note, however, that in the discrete-time case, if we use the simplified version (40) of the basic design problem then only the sufficient part of Theorem 9 is achieved.

We now generalize the results of Theorem 9 to uncertain discrete-time systems. To this end we assume that (E) is such that $F \in D_F$ where D_F is the convex domain given in (41).

We say that the collection of linear systems $\{(A_i, B_i), i = 1 \cdots M\}$ is quadratically stabilizable via linear control if there exist $W = W' > 0$ and K such that $(A_i - B_iK)W(A_i - B_iK)' - W < 0$ for all $i = 1 \cdots M$. Defining the convex cone

$$\mathscr{C}_U \triangleq \{\mathscr{W} = \mathscr{W}' \geqq 0 \in \Re^{p\times p}\colon \langle \mathscr{W}, Z\rangle < 0, Z \in \Sigma_U\}, \tag{63}$$

where $\Sigma_U \triangleq \Sigma_0 \cup \Sigma_1 \cup \cdots \cup \Sigma_M$ and Σ_i $i = 1 \cdots M$ are given by (58) with F replaced by F_i $i = 1 \cdots M$, we can prove that the collection $\{(A_i, B_i), i = 1 \cdots M\}$ is quadratically stabilizable via linear feedback control if and only if $\mathscr{C}_U \neq \varnothing$.

The proof of the main result of this section (Theorem 11) is based on the following property.

LEMMA 10. *Assume that $\mathscr{W} = \mathscr{W}' \geqq 0 \in \Re^{p\times p}$; then*

$$F\mathscr{W}F' \leqq \sum_{i=1}^{M} \lambda_i F_i \mathscr{W} F_i' \quad \forall F \in D_F. \tag{64}$$

Proof. For arbitrary but fixed $z \in \Re^p$ define the function $g(\cdot): \Re^{p\times p} \to \Re$

$$g(F) \triangleq z'F\mathscr{W}F'z. \tag{65}$$

This function is convex. Indeed, for all $F, \Gamma \in \Re^{p\times p}$ we have

$$\frac{d^2}{d\alpha^2} g(F+\alpha\Gamma) = 2z'\Gamma\mathscr{W}\Gamma'z \geqq 0, \tag{66}$$

where the inequality follows from the nonnegativeness of $\mathscr{W}$. Using this fact together with the definition of D_F, we get

$$\begin{aligned} z'F\mathscr{W}F'z &= g\left(\sum_{i=1}^{M} \lambda_i F_i\right) \\ &\leqq \sum_{i=1}^{M} \lambda_i g(F_i) \\ &= z'\left\{\sum_{i=1}^{M} \lambda_i F_i \mathscr{W} F_i'\right\} z. \end{aligned} \tag{67}$$

Since $z \in \Re^p$ is arbitrary, (67) implies (64) and the lemma is proved.

THEOREM 11. *Assume that $F \in D_F$. The system* (S) *is quadratically stabilizable via linear control if and only if the same holds for the collection* $\{(A_i, B_i), i = 1 \cdots M\}$.

Proof. The necessity is obvious. For the sufficiency, suppose the $\mathscr{C}_U \neq \emptyset$. Partitioning $\mathscr{W} \in \mathscr{C}_U$ according to (32), setting $W = W_1$, $K = W_2' W_1^{-1}$, and recalling that $\mathscr{W} \geqq 0$, we get for all $F \in D_F$ and $z' = [x' \;\; 0]$, $x \in \Re^n$ (see (62)):

$$x'[(A-BK)W(A-BK)' - W]x \leqq z'[F\mathscr{W}F' - \mathscr{W}]z. \tag{68}$$

Now, using (64) it follows that

$$\begin{aligned} z'[F\mathscr{W}F' - \mathscr{W}]z &\leqq z'\left[\sum_{i=1}^{M} \lambda_i F_i \mathscr{W} F_i' - \mathscr{W}\right] z \\ &\leqq \sum_{i=1}^{M} \lambda_i z'[F_i \mathscr{W} F_i' - \mathscr{W}]z \\ &< 0 \quad \forall F \in D_F. \end{aligned} \tag{69}$$

Finally, (69) together with (68) proves the theorem proposed.

This theorem implies that the algorithm proposed in § 5 can be still used to determine a small feedback gain just sufficient to stabilize a given uncertain discrete-time system. To this end it is sufficient to define $\Theta_i(\mathscr{W}) \triangleq F_i\mathscr{W}F_i' - \mathscr{W}$, $i = 1 \cdots M$. As before, this is a linear function of $\mathscr{W} \in \Re^{p\times p}$ and the relaxation procedure will determine the feedback gain whenever $\mathscr{C} \neq \emptyset$.

7. Examples. In this section, some examples covering the main problems discussed are given to illustrate the usefulness of the approach.

The first example is borrowed from [16], where it is shown how to stabilize a denumerable set of operating points (a set of linear systems) by means of a nonlinear control. The proposed approach has been run (results below) under the same data, resulting in a linear feedback gain stabilizing simultaneously each of the four systems.

The systems are of third order with the following structure:

$$\dot{x} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & -30 \end{bmatrix} x + \begin{bmatrix} b_1 \\ 0 \\ 30 \end{bmatrix} u$$

and the numerical data (in Table 1):

For the W_1 and W_2 matrices the final results are

$$W_1 = \begin{bmatrix} 171.5583 & -15.9527 & -5.1070 \\ -15.9527 & 2.5890 & 0.6497 \\ -5.1070 & 0.6497 & 1.5261 \end{bmatrix}, \qquad W_2 = \begin{bmatrix} 1.4005 \\ -0.7400 \\ 0.0444 \end{bmatrix}$$

resulting in a K such as

$$K = [-0.0411 \quad -0.5729 \quad 0.1355].$$

With the given approach, not only the four operating points are stabilized, but also all the systems given by a linear convex combination of the former. This is partly verified on Fig. 2 where the root loci of the following systems are plotted:

$$A = \mu(A_i - B_iK) + (1-\mu)(A_j - B_jK),$$
$$i \neq j, \quad i, j = 1, 2, 3, 4, \quad 0 \leqq \mu \leqq 1$$

in the open-loop ($K = 0$) and the closed-loop (K given above) cases. In open loop, some systems are unstable, with a part of the root locus inside the right half complex plane. It is verified that the closed-loop root loci all lie in the stable part of the complex plane.

Some experiments have also been performed for discrete-time systems. The one given below is derived from the preceding example in the following way:

$$x(t+1) = \tilde{A}_i x(t) + \tilde{B}_i u(t), \qquad i = 1, 2, 3, 4,$$

where

$$\tilde{A}_i = \exp[A_i \Delta t],$$
$$\tilde{B}_i = \int_0^{\Delta t} \exp(A_i \tau) B_i \, d\tau.$$

The sample period Δt has been chosen equal to 0.1, which gives the data (Table 2).

The same processing as for the continuous case was repeated, leading to the following results and the same conclusion as for the stability (see Fig. 3):

$$W_1 = \begin{bmatrix} 3620.22 & -335.211 & -122.406 \\ -335.211 & 50.9528 & 16.8288 \\ -122.406 & 16.8288 & 23.772 \end{bmatrix}, \qquad W_2 = \begin{bmatrix} 4.6330 \\ -6.0835 \\ 2.7064 \end{bmatrix}$$

resulting in a K such as

$$K = [-0.0230 \quad -0.3513 \quad 0.2441].$$

TABLE 1
Continuous operating points.

Operating point	1	2	3	4
a_{11}	−0.9896	−0.6607	−1.702	−0.5162
a_{12}	17.41	18.11	50.72	29.96
a_{13}	96.15	84.34	263.5	178.9
a_{21}	0.2648	0.08201	0.2201	−0.6896
a_{22}	−0.8512	−0.6587	−1.418	−1.225
a_{23}	−11.39	−10.81	−31.99	−30.38
b_1	−97.78	−272.2	−85.09	−175.6

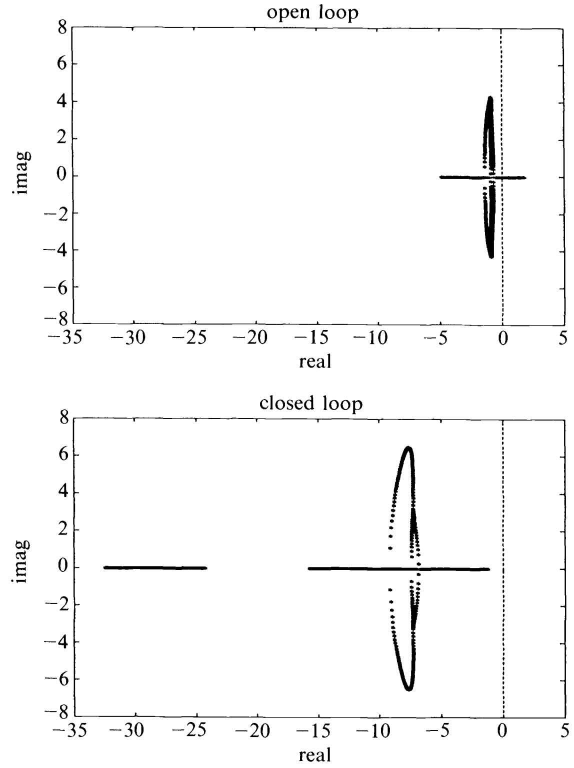


FIG. 2. *Open-loop and closed-loop continuous root loci.*

The last example concerns the problem of decentralization. It has been borrowed from [26], where it is presented as a mechanical manipulator control example. The difference here is twofold: First the decentralization constraint is imposed, then the uncertainty intervals on the parameters are enlarged.

The system is written as

$$\frac{d}{dt}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} -1/\tau_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ k_{11} & -k_{11} & (c_1 - k_{12}) & 0 & 0 & c_2 \\ 0 & 0 & 0 & -1/\tau_2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & c_3 & k_{21} & -k_{21} & -k_{22} \end{bmatrix}$$

$$\cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} + \begin{bmatrix} 1/\tau_1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1/\tau_2 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}.$$

The decentralized constraints restricts u_1 to involve (x_1, x_2, x_3) and u_2 to involve (x_4, x_5, x_6). The nominal system is defined by the following data:

$$k_{11} = k_{21} = 10, \qquad k_{12} = k_{22} = 2,$$

TABLE 2
Discrete operating points.

Operating point	1	2	3	4
$\tilde{a}_{11}$	0.9268	0.9430	0.8915	0.8648
$\tilde{a}_{12}$	1.6002	1.6996	4.4207	2.3959
$\tilde{a}_{13}$	2.4524	2.1347	4.3207	3.4747
$\tilde{a}_{21}$	0.0243	0.0077	0.0192	−0.0613
$\tilde{a}_{22}$	0.9396	0.9432	0.9162	0.8018
$\tilde{a}_{23}$	−0.2897	−0.3130	−0.8262	−1.0917
$\tilde{a}_{31}$	0	0	0	0
$\tilde{a}_{32}$	0	0	0	0
$\tilde{a}_{33}$	0.0498	0.0498	0.0498	0.0498
$\tilde{b}_1$	−3.5451	−21.2959	4.9973	−6.9694
$\tilde{b}_2$	−0.8124	−0.8095	−2.0311	−1.6916
$\tilde{b}_3$	0.9502	0.9502	0.9502	0.9502

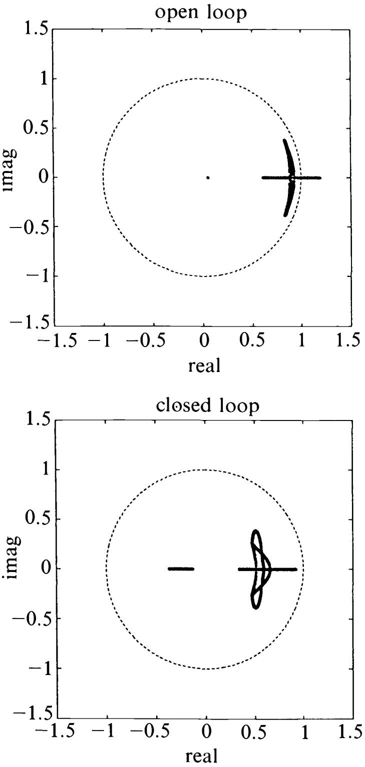


FIG. 3. *Open-loop and closed-loop discrete root loci.*

$$\tau_1 = \tau_2 = 0.1,$$

$$c_1 = 0.2, \qquad c_2 = c_3 = 0.1.$$

The first simulation (Fig. 4) is performed from the results in [26] on the nominal system, and the associated feedback K:

$$K = \begin{bmatrix} 5.67 & 114.0 & 14.4 & 0.017 & 0.321 & 0.094 \\ 0.017 & 0.321 & 0.094 & 5.63 & 113.0 & 14.2 \end{bmatrix}.$$

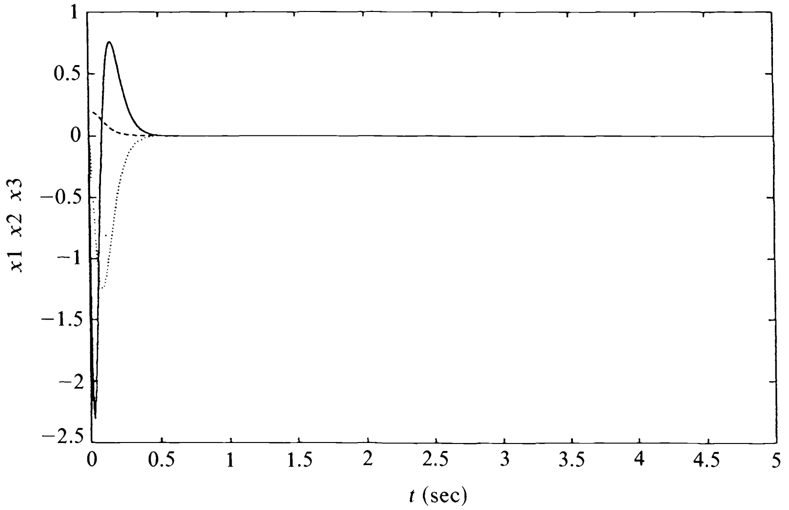


FIG. 4. *Nominal parameter values—centralized gain.*

Throughout the figures only the three first-state variables are plotted, from the initial conditions

$$x_0 = [0.3 \quad 0.2 \quad 0 \quad 0.7 \quad 0.2 \quad 0]'.$$

For the nominal system the proposed approach provided the W_1 and W_2 matrices

$$W_1 = \begin{bmatrix} 3.2437 & 0.8842 & -1.0682 & 0 & 0 & 0 \\ 0.8842 & 1.4231 & -0.7962 & 0 & 0 & 0 \\ -1.0682 & -0.7962 & 3.2437 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3.2437 & 0.8854 & -1.2115 \\ 0 & 0 & 0 & 0.8854 & 1.4657 & -0.9280 \\ 0 & 0 & 0 & -1.2115 & -0.9280 & 3.2437 \end{bmatrix},$$

$$W_2 = \begin{bmatrix} -0.8240 & 0 \\ -0.5724 & 0 \\ 1.5125 & 0 \\ 0 & -0.7607 \\ 0 & -0.6179 \\ 0 & 1.5161 \end{bmatrix}$$

from which follows the decentralized gain

$$K = \begin{bmatrix} -0.0869 & -0.1198 & 0.4083 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.0439 & -0.1337 & 0.4127 \end{bmatrix}.$$

Figure 5 shows the responses for the decentralized gain.

Finally, the case of both uncertainty and decentralized constraints has been run. The uncertainty affects the parameters τ_1, τ_2, c_1, c_2, and c_3 in the following way:

$$0.02 \leqq \tau_1 = \tau_2 \leqq 0.18,$$

$$-14.8 \leqq c_1 \leqq 15.2,$$

$$-7.9 \leqq c_2 = c_3 \leqq 8.1.$$

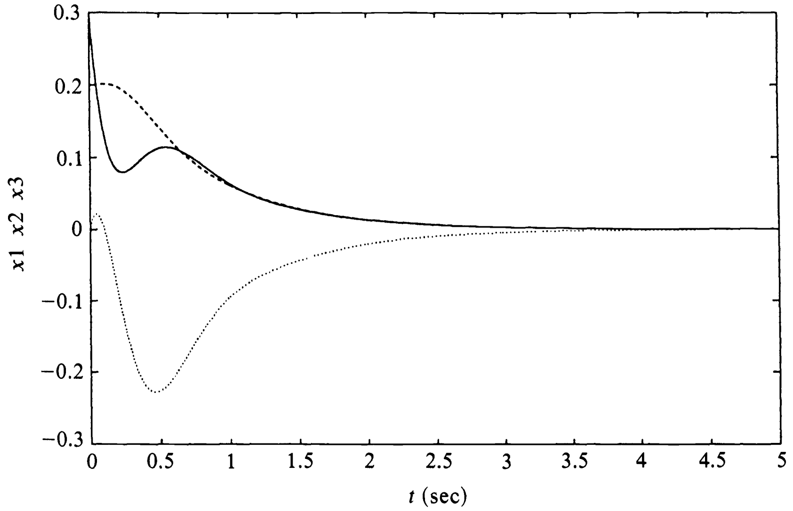


FIG. 5. *Nominal parameter values—decentralized gain.*

The numerical results are

$$W_1 = \begin{bmatrix} 38.4122 & 2.5394 & -14.175 & 0 & 0 & 0 \\ 2.5394 & 2.7764 & -0.2563 & 0 & 0 & 0 \\ -14.175 & -0.2563 & 7.7523 & 0 & 0 & 0 \\ 0 & 0 & 0 & 38.4122 & 1.9798 & -11.0742 \\ 0 & 0 & 0 & 1.9798 & 2.1106 & -0.1876 \\ 0 & 0 & 0 & -11.0742 & -0.1876 & 6.0045 \end{bmatrix},$$

$$W_2 = \begin{bmatrix} 21.2088 & 0 \\ -3.1099 & 0 \\ 19.6773 & 0 \\ 0 & 50.6388 \\ 0 & -1.7759 \\ 0 & 14.5368 \end{bmatrix},$$

which give the following robust gain:

$$K = \begin{bmatrix} 5.3936 & -4.9236 & 12.2377 & 0 & 0 & 0 \\ 0 & 0 & 0 & 4.6937 & -4.2712 & 10.9442 \end{bmatrix}.$$

In Figs. 6 and 7 are plotted the results of a numerical simulation where the coefficients are time-varying according to

$$\tau_1 = 0.1 + 0.08 \cos(0.1t),$$

$$\tau_2 = 0.1 + 0.08 \sin(0.5t),$$

$$c_1 = 0.2 + 15.0 \cos(0.1t),$$

$$c_2 = 0.1 - 8.0 \cos(0.2t),$$

$$c_3 = 0.1 - 8.0 \cos(0.1t).$$

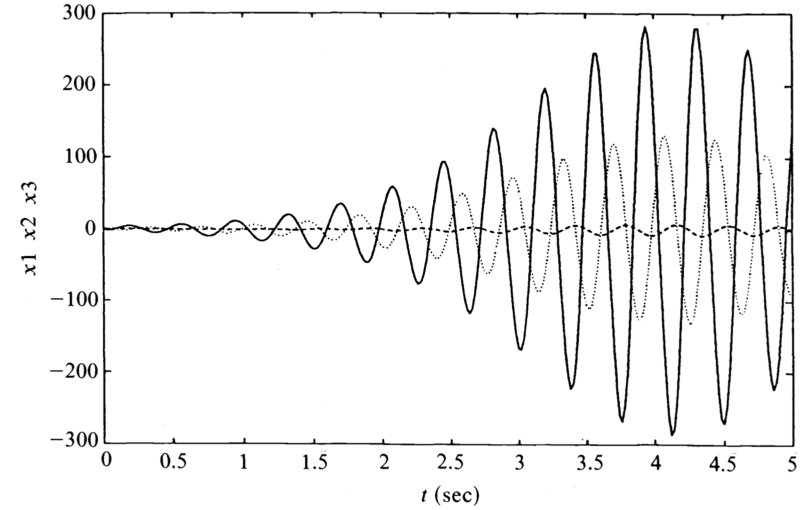


FIG. 6. *Uncertain parameters—centralized nominal gain.*

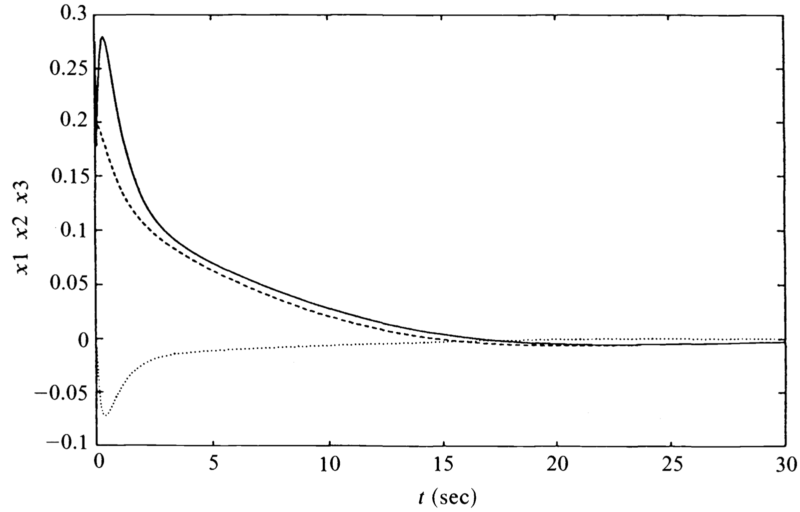


FIG. 7. *Uncertain parameters—decentralized robust gain.*

Figure 6 corresponds to the centralized gain given in [26], and Fig. 7 shows the one obtained by the robust decentralized gain given above. Note that the perturbations correspond to rather big variation around the nominal values. That almost causes divergence of the response plotted in Fig. 6 within a five-second interval time, while the plot of Fig. 7 drawn over a 20-second interval time shows a good damping of the response.

8. Conclusions. In this paper we introduced a new parameter space method for linear system control design. One of its most important features is that the feedback gain is determined from the solution of a *convex* problem. This is a key result because it opens the possibility of applying the most powerful techniques available to date in the mathematical programming literature for the solution of the above-mentioned problem. Furthermore, large-scale systems can be handled more efficiently by means of decomposition procedures (this important aspect has not been analysed in this paper).

The procedure is sufficiently general to solve several important design problems including the decentralized control problem and optimal control of uncertain systems by means of linear feedback. For them, necessary and sufficient conditions for solvability have been provided in terms of existence of some convex cones.

It is important to stress that the same algorithm can be used for both discrete-time and continuous-time systems control design. Indeed, the necessary and sufficient condition for the solvability of the quadratically stabilization problem of uncertain systems can be written as

$$\begin{gathered} T'\mathcal{W}T > 0, \\ T'\Theta_i(\mathcal{W})T < 0, \qquad i = 1 \cdots M, \\ \mathcal{W} \geqq 0, \end{gathered}$$

where $T \in \Re^{p\times n}$ is a matrix spanning the nullspace of G' and $\Theta_i(\cdot)$, $i = 1 \cdots M$ are linear functions of $\mathcal{W} \in \Re^{p\times p}$. At each iteration only the linear cut to be added to the master linear programming problem depends on the dynamic (continuous-time or discrete-time) representation of the system.

The present approach proposes a bridge between optimal control problems formulated on the parameter space and one of the most important and basic properties of mathematical programming programs—the convexity. As a natural consequence, it allows to solve several control problems in a unique and well-posed (numerically speaking) way, a fact which does not hold even for the LQ design.

Acknowledgment. The authors wish to thank Reviewer 2 for several helpful comments.

REFERENCES

[1] M. Athans, *The role and use of stochastic linear-quadratic-Gaussian problem in control systems design*, IEEE Trans. Automat. Control, 16 (1971), pp. 529-552.

[2] B. R. Barmish, *Stabilization of uncertain systems via linear control*, IEEE Trans. Automat. Control, 28 (1983), pp. 848-850.

[3] ———, *Necessary and sufficient conditions for quadratic stabilizability of an uncertain system*, J. Optim. Theory Appl., 46 (1985), pp. 399-408.

[4] J. Bernussou, P. L. D. Peres, and J. C. Geromel, *A linear programming oriented procedure for quadratic stabilization of uncertain systems*, Systems Control Lett., 13 (1989), pp. 65-72.

[5] ———, *Robust decentralized regulation: a linear programming approach*, IFAC/IFORS/IMACS Symposium-Large Scale Systems: Theory and Applications, Berlin—GDR, 1 (1989), pp. 135-138.

[6] S. P. Boyd, V. Balakrishnan, C. H. Barratt, N. M. Khraishi, X. Li, D. G. Meyer, and S. A. Norman, *A new* CAD *method and associated architectures for linear controllers*, IEEE Trans. Automat. Control, 33 (1988), pp. 268-283.

[7] J. C. Geromel, *On the determination of a diagonal solution of the Lyapunov equation*, IEEE Trans. Automat. Control, 30 (1985), pp. 404-406.

[8] J. C. Geromel and J. Bernussou, *Optimal decentralized control of dynamic systems*, Automatica, 18 (1982), pp. 545-557.

[9] J. C. Geromel and J. J. da Cruz, *On the robustness of optimal regulators for nonlinear discrete-time systems*, IEEE Trans. Automat. Control, 32 (1987), pp. 703-710.

[10] J. C. Geromel and A. Yamakami, *Stabilization of continuous and discrete linear systems subjected to control structure constraints*, Internat. J. Control, 63 (1982), pp. 429-444.

[11] H. P. Horisberger and P. R. Bélanger, *Regulators for linear time invariant plants with uncertain parameters*, IEEE Trans. Automat. Control, 21 (1976), pp. 705-708.

[12] P. P. KHARGONEKAR AND M. A. ROTEA, *Stabilization of uncertain systems with norm bounded uncertainty using control Lyapunov functions*, in Proc. 27th Conference on Decision and Control, Austin, TX, December 1988, pp. 503-507A.

[13] L. S. LASDON, *Optimization Theory for Large Scale Systems*, MacMillan, London, 1972.

[14] D. G. LUENBERGER, *Introduction to Linear Programming*, Addison-Wesley, Reading, MA, 1973.

[15] P. M. MÄKILÄ AND H. T. TOIVONEN, *Computational methods for parametric* LQ *problems—A survey*, IEEE Trans. Automat. Control, 32 (1987), pp. 658-671.

[16] I. R. PETERSEN, *A procedure for simultaneously stabilizing a collection of single input linear systems using non-linear state feedback control*, Automatica, 23 (1987), pp. 33-40.

[17] I. R. PETERSEN AND C. V. HOLLOT, *A Riccati equation approach to the stabilization of uncertain linear systems*, Automatica, 22 (1986), pp. 397-411.

[18] E. POLAK, *On the mathematical foundations of nondifferentiable optimization in engineering design*, SIAM Rev., 29 (1987), pp. 21-89.

[19] E. POLAK AND D. M. STIMLER, *Majorization: a computational complexity reduction technique in control system design*, IEE Trans. Automat. Control, 33 (1988), pp. 1010-1021.

(20) E. POLAK AND S. E. SALCUDEAN, *On the design of linear multivariable feedback systems via constrained nondifferentiable optimization in H^∞ spaces*, IEEE Trans. Automat. Control, 34 (1989), pp. 268-276.

[21] M. G. SAFONOV AND M. ATHANS, *Gain and phase margins for multiloop* LQG *regulators*, IEEE Trans. Automat. Control, 22 (1987), pp. 173-179.

[22] M. G. SAFONOV, *Stability and Robustness of Multivariable Feedback Systems*, M.I.T. Press, Cambridge, MA, 1980.

[23] D. D. SILJAK, *Large Scale Dynamic Systems: Stability and Structure*, North-Holland, New York, 1978.

[24] ———, *Parameter space methods for robust control design: a guided tour*, Tech. Report EECS-031588, School of Engineering, Santa Clara University, Santa Clara, CA, January 1988.

[25] W. M. WONHAM, *Linear Multivariable Control*, Lecture Notes in Economics and Mathematical Systems, Vol. 101, Springer-Verlag, Berlin, New York, 1974.

[26] M. A. ZOHDY, N. K. LOH, AND A. A. ABDUL-WAHAB, *A robust optimal model matching control*, IEEE Trans. Automat. Control, 32 (1987), pp. 410-414.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 403-419, March 1991

ON THE CONVERGENCE OF THE PROXIMAL POINT ALGORITHM FOR CONVEX MINIMIZATION*

OSMAN GÜLER†

Abstract. The proximal point algorithm (PPA) for the convex minimization problem $\min_{x\in H} f(x)$, where $f: H \to R \cup \{\infty\}$ is a proper, lower semicontinuous (lsc) function in a Hilbert space H is considered. Under this minimal assumption on f, it is proved that the PPA, with positive parameters $\{\lambda_k\}_{k=1}^{\infty}$, converges in general if and only if $\sigma_n = \sum_{k=1}^{n} \lambda_k \to \infty$. Global convergence rate estimates for the residual $f(x_n) - f(u)$, where x_n is the nth iterate of the PPA and $u \in H$ is arbitrary are given. An open question of Rockafellar is settled by giving an example of a PPA for which x_n converges weakly but not strongly to a minimizer of f.

Key words. proximal point algorithm, convex programming, strong convergence

AMS(MOS) subject classifications. primary 90C25; secondary 49D45, 49D37

1. Introduction.

Let H be a real Hilbert space. We consider the minimization problem

$$\min_{x\in H} f(x), \tag{1.1}$$

where $f: H \to \mathbf{R} \cup \{\infty\}$ is a proper, lower semicontinuous (lsc) convex function, where we follow the terminology established in Aubin and Ekeland [1] or Rockafellar [16]. Many convex programming problems with or without constraints can be reduced to (1.1).

One method for solving (1.1) is the *proximal point algorithm* (PPA) first introduced by Martinet [10]. The PPA is based on the notion of *proximal mapping* J_λ,

$$J_\lambda(x) = x_\lambda = \arg\min_{x\in H} \left\{ f(z) + \frac{1}{2\lambda} \|z - x\|^2 \right\}, \tag{1.2}$$

introduced by Moreau [12]. The PPA is an iterative procedure, which starts at a point $x_0 \in H$, and generates recursively a sequence of points $x_{k+1} = J_{\lambda_{k+1}(x_k)}$, where $\{\lambda_k\}_{k=1}^{\infty}$ is a sequence of positive numbers.

It turns out that a proximal mapping can be defined for an arbitrary *maximal monotone operator* $A: H \to H$. Recall that a multivalued mapping $A: H \to H$ is said to be a *monotone operator* if $w' \in A(x')$ and $w \in A(x)$ imply $\langle w' - w, x' - x \rangle \geqq 0$. Clearly, if A is a monotone operator, then

$$\begin{aligned} w \in A(x),\ w' \in A(x') \Rightarrow & \|(x' + w') - (x + w)\|^2 \\ & \geqq \|x' - x\|^2 + \|w' - w\|^2; \end{aligned} \tag{1.3}$$

in particular

$$x' \neq x \Rightarrow (I + A)(x) \cap (I + A)(x') = \varnothing. \tag{1.4}$$

A monotone operator A is said to be *maximal monotone* if the graph $G(A) = \{(w, x) \in H \times H \mid w \in A(x)\}$ is not properly contained in the graph of any other monotone operator $A': H \to H$. A *solution* to A is a point $x^* \in H$ such that $0 \in A(x^*)$.

Many problems that involve convexity can be formulated as finding the solution of a maximal monotone operator. For example, convex minimization, concave-convex

* Received by the editors July 17, 1989; accepted for publication (in revised form) March 2, 1990.

† Department of Management Sciences, University of Iowa, Iowa City, Iowa 52242.

saddle-point problems, and solutions of games can be formulated in this way. In particular, the subdifferential $A=\partial f$ is a maximal monotone operator, and a point $x^*\in H$ minimizes f if and only if $0\in\partial f(x^*)$. The classical result of Minty [11] states that a monotone operator A is maximal if and only if $I+A$ is surjective. If A is a maximal monotone operator and $\lambda>0$, the operator J_λ, defined by $J_\lambda(x)=(I+\lambda A)^{-1}(x)$, is called the *resolvent* of A. It follows from (1.4) that the resolvent J_λ is a single-valued operator on H. Moreover, (1.3) implies that J_λ is nonexpansive: that is, if $x, y\in H$, $\|J_\lambda(y)-J_\lambda(x)\|\leqq\|y-x\|$. Also, the *Yosida approximation* A_λ, $A_\lambda(x)=(x-J_\lambda(x))/\lambda$, is Lipschitz continuous with constant $1/\lambda$. That is, for $x, y\in H$, $\|A_\lambda(y)-A_\lambda(x)\|\leqq\|y-x\|/\lambda$.

The PPA for a maximal monotone operator is an iterative procedure that starts at a point $x_0\in H$, and generates recursively a sequence of points $x_{k+1}=J_{\lambda_{k+1}}(x_k)$, where $\{\lambda_k\}_{k=1}^\infty$ is a sequence of positive numbers. It is treated in the papers [4], [7], [10], [17], and [18]. The important paper of Brézis and Lions [4] contains many interesting results. Rockafellar [18] shows how the PPA can be applied in convex programming. We stress that when $A=\partial f$, the PPA described here reduces to the iteration described above in the context of (1.1).

Notation. We use the following notation in the paper. If f is a proper, lsc convex function on H, the *effective domain* of f is the set $\{x\in H: f(x)<\infty\}$, which we denote by $D(f)$. We will sometimes refer to lsc convex functions as *closed* functions. The infimum of f is denoted by $f^*=\inf_{x\in H}f(x)$, and the set of minimizers of f (possibly empty) is denoted by $X^*=\{x\in H: f(x)=f^*\}$. If $A: H\to H$ is a multivalued operator, the *domain* of A is the set $D(A)=\{x\in H: A(x)\neq\emptyset\}$, and the *range* of A is the set $R(A)=\bigcup\{A(x): x\in D(A)\}$. If the sequence $\{\lambda_k\}_{k=1}^\infty$ of positive numbers lists the proximal parameters, we define $\sigma_n=\sum_{k=1}^n\lambda_k$. By convention $\sigma_0=0$. If the sequence $\{x_k\}_{k=0}^\infty$ is the trajectory of a PPA, we will write $y_k\equiv A_{\lambda_k}(x_{k-1})=(x_{k-1}-x_k)/\lambda_k$. We use $J_\lambda(x)$ and x_λ interchangeably. If A is maximal monotone then $A(x)$ is closed and convex (see Aubin and Ekeland [1, Prop. 3, §6.7]. In this case, if $A(x)\neq\emptyset$, we denote the least norm element of $A(x)$ by A^0x. For any set $S\subseteq H$, we define the *distance function* $\rho(x, S)=\inf\{\|x-s\|: s\in S\}$.

Every maximal monotone operator engenders a *nonlinear contractive semigroup* $\{S(t): t\geqq 0\}$ of maps $S(t):\overline{D(A)}\to\overline{D(A)}$, satisfying the following properties for $t, s\geqq 0$ and $x, y\in\overline{D(A)}$:

(i) $S(0)x=x$,

(ii) $S(t+s)x=S(t)S(s)x$ (semigroup property), and

(iii) $\|S(t)x-S(t)y\|\leqq\|x-y\|$.

Indeed, $S(t)x=u(t)$, where $u(t)$ is the unique solution to the *differential inclusion*

$$\frac{du}{dt}\in -Au(t),\qquad u(0)=x. \tag{1.5}$$

For an excellent treatment of nonlinear contractive semigroups in a Hilbert space, the reader is referred to Brézis [3].

There is an intimate relationship between nonlinear (contractive) semigroups and the proximal point algorithm. If we discretize the differential inclusion (1.5) by the backward Euler differencing, we obtain

$$\frac{x_k-x_{k-1}}{\lambda_k}\in -A(x_k), \tag{1.6}$$

and we obtain $x_k=(I+\lambda_kA)^{-1}x_{k-1}\equiv J_{\lambda_k}(x_{k-1})$. Therefore, PPA is just the backward Euler discretization of the differential inclusion (1.5). It is important to keep this connection in mind, since PPA inherits many of the nice properties of the contractive semigroup $S(t)$ and vice versa. See §5 for details.

In this paper, we restrict our attention to the case $A = \partial f$ for two reasons. The first reason is subdifferentials of convex functions form an important subclass of maximal monotone operators. The second, and perhaps the more important reason, is that the operator ∂f has special properties (for example, demipositivity; see Bruck [5]) not shared by other maximal monotone operators. We exploit the special properties of ∂f to obtain sharper results.

In the literature, the convergence properties of the PPA are studied only in the case where f has a minimizer, and the convergence rate of the algorithm is given only in the case where f is strongly convex. Moreover, the convergence rate is given in terms of the closeness of x_k to a minimizer of f. We depart from this tradition. We give convergence of the PPA under the weakest conditions, even in cases where f has no minimizer, or is unbounded from below. Our convergence rate results are in terms of the residual $f(x_k) - f(u)$ where u is an arbitrary point in H.

The organization of the rest of the paper is as follows. In §2, we establish the convergence properties of the PPA under the weakest possible assumptions. We establish global convergence rate results along with some interesting results which we use in later sections. In §3, we sharpen the convergence rate result for the residual $f(x_n) - f^*$ in the case where the PPA trajectory converges strongly to a minimizer of f. In §4, we present a fundamental estimate due to Kobayashi. In §5, we answer an open question posed by Rockafellar [17]: Does the PPA always converge strongly? We give a proper, closed function in an infinite-dimensional Hilbert space for which the PPA converges weakly but not strongly.

2. The convergence of the proximal point algorithm. Let H be a Hilbert space and $f: H \to \mathbf{R} \cup \{\infty\}$ be a proper, closed convex function. We are concerned with the convergence properties of the PPA applied to the minimization of f. In the literature, convergence results for the PPA are given only in the case where f has a minimizer, and convergence rate results are given in the case in which f enjoys strong convexity properties. Moreover, the convergence rate results are only asymptotic.

In this section, we prove the convergence of the PPA under the weakest possible conditions and provide global convergence rate estimates for the residual $f(x_n) - f(u)$, where x_n is the nth iterate of the PPA and u is any point in H. The behavior of x_n and $y_n = (x_{n-1} - x_n)/\lambda_n$ is also studied.

The following result is well known. Since the proof is short, we include it.

LEMMA 2.1. *$\{\|y_n\|\}_{n=1}^{\infty}$ is a decreasing sequence.*

Proof. Since $y_n \in \partial f(x_n)$, $y_{n+1} \in \partial f(x_{n+1})$, and ∂f is a monotone operator, we have $\langle y_{n+1} - y_n, x_{n+1} - x_n \rangle \geqq 0$. Since $y_{n+1} = (x_n - x_{n+1})/\lambda_{n+1}$, we obtain $\langle y_{n+1} - y_n, y_{n+1} \rangle \leqq 0$, which implies $\|y_{n+1}\|^2 \leqq \langle y_n, y_{n+1} \rangle \leqq \|y_n\| \cdot \|y_{n+1}\|$. The lemma is proved. □

The following result contains the *fundamental estimate* from which we derive most of the convergence results of this section.

LEMMA 2.2. *Let $\{\lambda_j\}_{j=1}^{\infty}$ be an arbitrary sequence of positive numbers. Suppose the* PPA *starts at x_0 and generates the sequence $\{x_n\}_{n=0}^{\infty}$, where $x_n = J_{\lambda_n}(x_{n-1})$. Then for any $u \in H$,*

$$f(x_n) - f(u) \leqq \frac{\|u - x_0\|^2}{2\sigma_n} - \frac{\|u - x_n\|^2}{2\sigma_n} - \frac{\sigma_n}{2} \|y_n\|^2. \tag{2.1}$$

Proof. Recall that $y_k = (x_{k-1} - x_k)/\lambda_k \in \partial f(x_k)$. By the convexity of f we have

$$f(u) - f(x_k) \geqq \langle y_k, u - x_k \rangle = \lambda_k^{-1} \langle x_{k-1} - x_k, u - x_k \rangle. \tag{2.2}$$

Therefore,

$$\begin{aligned}2\lambda_k(f(u)-f(x_k)) &\geqq 2\langle x_{k-1}-x_k, u-x_k\rangle \\ &= \|x_{k-1}-x_k\|^2+\|u-x_k\|^2-\|u-x_{k-1}\|^2 \\ &= \lambda_k^2\|y_k\|^2+\|u-x_k\|^2-\|u-x_{k-1}\|^2.\end{aligned} \tag{2.3}$$

Summing (2.3) for $k=1,\cdots,n$, we obtain

$$2\sigma_n f(u)-2\sum_{k=1}^n \lambda_k f(x_k) \geqq \sum_{k=1}^n \lambda_k^2\|y_k\|^2+\|u-x_n\|^2-\|u-x_0\|^2. \tag{2.4}$$

Setting x_{k-1} for u in (2.2) yields

$$f(x_{k-1})-f(x_k) \geqq \lambda_k^{-1}\|x_{k-1}-x_k\|^2=\lambda_k\|y_k\|^2. \tag{2.5}$$

Recall that $\sigma_k=\sum_{j=1}^k \lambda_j$, for $k\geqq 1$. Multiplying (2.5) by σ_{k-1}, we obtain

$$\sigma_{k-1}f(x_{k-1})-\sigma_k f(x_k)+\lambda_k f(x_k) \geqq \sigma_{k-1}\lambda_k\|y_k\|^2.$$

Summing the last inequality for $k=1,\cdots,n$ and noting $\sigma_0=0$, we obtain

$$-\sigma_n f(x_n)+\sum_{k=1}^n \lambda_k f(x_k) \geqq \sum_{k=2}^n \sigma_{k-1}\lambda_k\|y_k\|^2. \tag{2.6}$$

Adding twice (2.6) to (2.4) yields

$$\begin{aligned}2\sigma_n(f(u)-f(x_n)) &\geqq 2\sum_{k=2}^n \sigma_{k-1}\lambda_k\|y_k\|^2+\sum_{k=1}^n \lambda_k^2\|y_k\|^2+\|u-x_n\|^2-\|u-x_0\|^2 \\ &\geqq \left(\sum_{k=1}^n \lambda_k^2+2\sum_{k=2}^n \sigma_{k-1}\lambda_k\right)\|y_n\|^2+\|u-x_n\|^2-\|u-x_0\|^2 \\ &= \sigma_n^2\|y_n\|^2+\|u-x_n\|^2-\|u-x_0\|^2,\end{aligned}$$

where the second inequality follows from Lemma 2.1. Rearranging the terms of the inequality above gives (2.1). □

The next theorem contains the convergence properties of the PPA under the weakest possible assumptions. It is the main result of this section.

THEOREM 2.1. *Let the sequence $\{x_n\}_{n=0}^\infty$ be the trajectory of a* PPA. *For any $u\in H$ the following global convergence estimate holds*:

$$f(x_n)-f(u) \leqq \frac{\|u-x_0\|^2}{2\sigma_n}. \tag{2.7}$$

Consequently, if $\sigma_n\to\infty$, then $f(x_n)\downarrow f^*=\inf_{x\in H}f(z)$. If $X^*\neq\emptyset$, then x_n converges weakly to a minimizer of f. Moreover,

$$f(x_n)-f^* \leqq \frac{\rho(x_0,X^*)^2}{2\sigma_n}. \tag{2.8}$$

Proof. The estimate (2.7) follows immediately from (2.1). In order to prove that $f(x_n)$ converges to f^*, we first consider the case $f^*>-\infty$. Let $\varepsilon>0$ be arbitrary, and choose a point x^ε such that $f(x^\varepsilon)\leqq f^*+\varepsilon$. From (2.7) we obtain $f(x_n)\leqq f^*+\varepsilon+\|x^\varepsilon-x\|^2/(2\sigma_n)$. Since $\sigma_n\to\infty$ as $n\to\infty$, we have $f(x_n)\leqq f^*+2\varepsilon$ for large enough n. Line (2.5) shows that $f(x_n)$ is nonincreasing. Since ε is arbitrary, $f(x_n)\downarrow f^*$. The proof of the convergence in the case $f^*=-\infty$ is similar.

It remains to prove the assertions about the case $X^* \neq \emptyset$. In this case, the weak convergence of x_n to a minimizer of f is proved in Brézis and Lions [4, Thm. 9]. Formula (2.8) follows by substituting x^* for u in (2.7), where x^* is the point in X^* closest to x. □

Remark 2.1. The condition $\sigma_n \to \infty$ is the *weakest* condition in order to ensure that $f(x_n) \downarrow f^*$. If $\sigma_n \to \sigma < \infty$, then x_n always *converges strongly*:

$$(2.9) \qquad \|x_{n+p} - x_n\| \leqq \sum_{j=n+1}^{n+p} \|x_{j-1} - x_j\| = \sum_{j=n+1}^{n+p} \lambda_j \|y_j\| \leqq \left(\sum_{j=n+1}^{n+p} \lambda_j \right) \|y_{n+1}\|.$$

Since $\sigma_n \to \sigma$, (2.9) shows that x_n is a Cauchy sequence, and therefore converges strongly to some point x^∞, even if f does not have a minimizer! Even if $X^* \neq \emptyset$, we have

$$\|x - x^\infty\| \leqq \sum_{j=1}^{\infty} \|x_{j-1} - x_j\| = \sum_{j=1}^{\infty} \lambda_j \|y_j\| \leqq \sigma \|y_1\|,$$

so that $\rho(x^\infty, X^*) \geqq \rho(x, X^*) - \|x - x^\infty\| \geqq \rho(x, X^*) - \sigma\|y_1\|$. If σ is small, then $\rho(x^\infty, X^*) > 0$, and $x^\infty \notin X^*$.

Remark 2.2. In [7], Güler introduced new proximal point algorithms for minimizing f. The first of these algorithms converges under the condition $\sum_{k=1}^{\infty} \lambda_k^{1/2} = \infty$. Note that under this condition, which is weaker than $\sigma_n \to \infty$, the standard PPA need not converge.

By setting x_0 for u in (2.1), we obtain the following result.

COROLLARY 2.1. *In a proximal point algorithm the following estimate holds*:

$$(2.10) \qquad f(x_n) \leqq f(x_0) - \frac{\|x_0 - x_n\|^2}{2\sigma_n}.$$

The next result will be useful in this section as well as in §5.

THEOREM 2.2. *Let* $A = \partial f$ *and* $u \in D(A)$. *In a proximal point algorithm the following estimates hold*:

$$(2.11) \qquad \|y_n\| \leqq \frac{\|x_n - x_0\|}{\sigma_n},$$

$$(2.12) \qquad \|y_n\| \leqq \|A^0 u\| + \frac{\|u - x_0\|}{\sigma_n}.$$

Proof. Formula (2.11) follows by substituting x_n for u in (2.1). From the convexity of f we obtain $f(x_n) \geqq f(u) + \langle A^0 u, x_n - u \rangle$, which implies $f(u) - f(x_n) \leqq \|A^0 u\| \cdot \|x_n - u\|$. Using this inequality in (2.1), we obtain

$$\begin{aligned} \sigma_n^2 \|y_n\|^2 &\leqq \|u - x_0\|^2 - \|u - x_n\|^2 + 2\sigma_n \|A^0 u\| \cdot \|x_n - u\| \\ &\leqq \|u - x_0\|^2 - \|u - x_n\|^2 + (\sigma_n^2 \|A^0 u\|^2 + \|x_n - u\|^2) \\ &= \|u - x_0\|^2 + \sigma_0^2 \|A^0 u\|^2 \\ &\leqq (\|u - x_0\| + \sigma_n \|A^0 u\|)^2. \end{aligned}$$

The theorem is proved. □

Remark 2.3. In Theorem 9 of [4] Brézis and Lions prove a weaker version of (2.12), in a special case. In particular, they prove that if $X^* \neq \emptyset$, then $\|y_n\| \leqq \sqrt{2}\, \rho(x, X^*)/\sigma_n$. However, their proof can be modified along the lines of the proof of our Lemma 2.2 so as to eliminate the factor $\sqrt{2}$. Our stronger estimate (2.12) is essential to prove (i) of Theorem 2.3 below.

The following continuous version of Theorem 2.2 can be obtained from it by passing to the limit. It will be needed in §5.

COROLLARY 2.2. *Let $A = \partial f$. For any $x \in \overline{D(A)}$ and $u \in D(A)$, we have*

$$\|A^0 S(t)x\| \leqq \frac{\|S(t)x - x\|}{t}, \tag{2.13}$$

$$\|A^0 S(t)x\| \leqq \|A^0 u\| + \frac{\|u - x\|}{t}. \tag{2.14}$$

Remark 2.4. A different proof of Corollary 2.2 is given in Brézis [3, Thm. 2.3.2].

The following result gives information about the behavior of x_n and y_n in a PPA.

THEOREM 2.3. *Let $A = \partial f$ and let v be the least norm element of $\overline{R(\partial f)}$. If $\sigma_n \to \infty$, then*:

(i) *y_n converges strongly to v,*

(ii) *x_n/σ_n converges strongly to $-v$.*

(iii) *$\{x_n\}_{n=0}^{\infty}$ is bounded if and only if f has a minimizer, that is, $X^* \neq \varnothing$. We have $\|x_n\| \to \infty$ if and only if $X^* = \varnothing$.*

Proof. $\overline{R(\partial f)}$ is a closed convex set. See, for example, Brézis [3, Thm. 2.2.2]. Therefore, the least norm element, $v \in \overline{R(\partial f)}$ exists and is the projection of zero onto $\overline{R(\partial f)}$. Let $\varepsilon > 0$ be arbitrary, and choose $x^\varepsilon \in D(\partial f)$ such that $\|A^0 x^\varepsilon\| \leqq v\| + \varepsilon$. Substituting x^ε for u in (2.12), and letting $n \to \infty$, we obtain

$$\lim_{n\to\infty} \|y_n\| = \inf_{n \geqq 1} \|y_n\| \leqq \|A^0 x^\varepsilon\| \leqq \|v\| + \varepsilon,$$

where we use the fact that $\sigma_n \to \infty$ in the first inequality above. Since ε is arbitrary, $\lim_{n\to\infty} \|y_n\| = \|v\|$. By the parallelogram identity, $\|y_n - v\|^2 + \|y_n + v\|^2 = 2\|y_n\|^2 + 2\|v\|^2$, which implies

$$\|y_n - v\|^2 = 2\|y_n\|^2 + 2\|v\|^2 - 4\|(y_n + v)/2\|^2. \tag{2.15}$$

Now, since $\overline{R(\partial f)}$ is convex, $(y_n + v)/2 \in \overline{R(\partial f)}$, and thus we have $\|(y_n + v)/2\| \geqq \|v\|$. By triangular inequality we also have $\|(y_n + v)/2\| \leqq \|y_n\|/2 + \|v\|/2 \to \|v\|$ as $n \to \infty$. Therefore $\|(y_n + v)/2\| \to \|v\|$. Letting $n \to \infty$ in (2.15), we conclude that y_n converges strongly to v and (i) is proved.

To prove (ii), we note that $(x - x_n)/\sigma_n = \sigma_n^{-1} \sum_{i=1}^n (x_{i-1} - x_i) = \sigma_n^{-1} \sum_{i=1}^n \lambda_i y_i$. Since $y_n \to v$ strongly, by an application of the Silverman-Toeplitz theorem (see, for example, Dunford and Schwartz [6]), we obtain that $(x - x_n)/\sigma_n$ converges strongly to v. Since $\sigma_n \to \infty$, x_n/σ_n converges strongly to $-v$. This proves (ii). Part (iii) is known in the literature, See, for example, Reich [15] for a proof. □

Remark 2.5. Reich [15] actually proves (ii) for an arbitrary maximal monotone operator by a different method.

COROLLARY 2.3. *Let $A = \partial f$. Suppose a* PPA *generates the sequence $\{x_n\}_{n=0}^{\infty}$, where $\sigma_n = \sum_{k=1}^n \lambda_k \to \infty$. Then, if $f^* > -\infty$, y_n converges strongly to zero. Consequently, if $f^* > -\infty$, then $0 \in \overline{R(\partial f)}$.*

Proof. Summing (2.5) for $k = 1, \cdots, n$, we obtain

$$\infty > f(x) - f^* \geqq f(x) - f(x_n) \geqq \sigma_n \|y_n\|^2.$$

Since $\sigma_n \to \infty$, we have $y_n \to 0 = v$, where v is the least norm element of $\overline{R(\partial f)}$. □

Remark 2.6. Ekeland's ε-variational principle (see Aubin and Ekeland [1, Chap. 5]) can be used to prove the fact that if $f^* > -\infty$, then there exists x_n, y_n with $y_n \in \partial f(x_n)$, $f(x_n) \downarrow f^*$, and $y_n \to 0$. Corollary 2.3 shows that such x_n and y_n can be generated by a PPA.

Remark 2.7. It is tempting to conjecture the converse of Corollary 2.3, namely, if $0 \in \overline{R(\partial f)}$, then $f^* > -\infty$. However, this conjecture is false. In the first draft of the paper we had a counterexample. One of the referees suggested the following simple counterexample:

$$f(x) = \begin{cases} -\sqrt{x} & \text{if } x \geqq 0, \\ +\infty & \text{otherwise.} \end{cases}$$

Observe that $f(x) \to -\infty$ as $x \to \infty$, while $\partial f(x) \to 0$. The other referee suggested another simple counterexample:

$$f(x) = \begin{cases} 1 - x & \text{if } x \leqq 1, \\ -\log(x) & \text{if } x \geqq 1. \end{cases}$$

Again $f(x) \to -\infty$ as $x \to \infty$, while $\partial f(x) \to 0$.

3. The convergence rate of the proximal point algorithm. Let $f: H \to \mathbf{R} \cup \{\infty\}$ be a proper, closed convex function. Assume that $X^* \neq \emptyset$, that is, f has minimizers. Let $\{\lambda_j\}_{j=1}^{\infty}$ be a sequence of positive numbers with $\sigma_n \to \infty$. Consider the proximal point algorithm for minimizing f, with parameters $\{\lambda_k\}$, starting at an initial point $x_0 \in H$. We saw in §2 that the points x_k generated by the PPA converge weakly to a minimizer of f. Using Theorem 2.1, we have $f(x_n) - f^* \leqq \rho(x, X^*)^2/(2\sigma_n)$, which implies $f(x_n) - f^* = O(\sigma_n^{-1})$. We shall see in §5 that x_n need not converge strongly to any minimizer of f. However, if x_n *does* converge strongly to a minimizer of f, we can improve the converge rate $O(\sigma_n^{-1})$ above to $o(\sigma_n^{-1})$.

THEOREM 3.1. *Let $\{\lambda_n\}_{n=1}^{\infty}$ be a sequence of positive numbers and $\sigma_n \to \infty$. Let $f: H \to \mathbf{R} \cup \{\infty\}$ be a proper, closed convex function which has a minimizer. Consider the* PPA *starting at $x = x_0$ and generating the points $x_n = (x_{n-1})_{\lambda_n}$. If x_n converges strongly to a minimizer of f, then the convergence rate estimate*

$$f(x_n) - f^* = o\left(\frac{1}{\sigma_n}\right),$$

holds, that is, $\sigma_n(f(x_n) - f^*) \to 0$.

Proof. Suppose x_n converges strongly to $x^* \in X^*$. For brevity, we denote that $W_k = f(x_k) - f(x^*) = f(x_k) - f^*$. We can rewrite (2.5) as

$$W_{k-1} - W_k \geqq \lambda_k^{-1} \|x_{k-1} - x_k\|^2. \tag{3.1}$$

Substituting x^* for u in (2.2), we obtain

$$\begin{aligned} f(x^*) &\geqq f(x_k) + \lambda_k^{-1}\langle x_{k-1} - x_k, x^* - x_k\rangle \\ &= f(x_k) + \lambda_k^{-1}\langle x_{k-1} - x_k, x^* - x_{k-1}\rangle + \lambda_k^{-1}\|x_{k-1} - x_k\|^2 \\ &\geqq f(x_k) - \lambda_k^{-1}\|x_{k-1} - x_k\| \cdot \|x_{k-1} - x^*\|. \end{aligned}$$

Therefore, $\|x_{k-1} - x_k\| \geqq \lambda_k W_k \|x_{k-1} - x^*\|^{-1}$. Using this inequality in (3.1), we obtain

$$\begin{aligned} W_{k-1} &\geqq W_k + \frac{1}{\lambda_k}(\lambda_k W_k \|x_{k-1} - x^*\|^{-1})^2 \\ &= W_k + \frac{\lambda_k}{\|x_{k-1} - x^*\|^2} W_k^2 = W_k\left(1 + \frac{\lambda_k}{\|x_{k-1} - x^*\|^2} W_k\right). \end{aligned}$$

Inverting this inequality, we obtain

$$W_{k-1}^{-1} \leqq W_k^{-1}\left(1 + \frac{\lambda_k}{\|x_{k-1} - x^*\|^2} W_k\right)^{-1}. \tag{3.2}$$

We want to obtain a recursive inequality from (3.2). We have from (1.2),

$$f(x_k) \leqq f(x_k) + \frac{1}{2\lambda_k}\|x_k - x_{k-1}\|^2 \leqq f(x^*) + \frac{1}{2\lambda_k}\|x^* - x_{k-1}\|^2,$$

so that $0 \leqq W_k\lambda_k\|x^* - x_{k-1}\|^{-2} \leqq \frac{1}{2}$. The function $(1+t)^{-1}$ is convex for $t > -1$, hence $(1+t)^{-1} \leqq 1 - 2t/3$, for $t \in [0, \frac{1}{2}]$. Using this fact, we obtain from (3.2) that

$$W_{k-1}^{-1} \leqq W_k^{-1}\left(1 - \frac{2\lambda_k}{3\|x_{k-1} - x^*\|^2} W_k\right) = W_k^{-1} - \frac{2\lambda_k}{3\|x_{k-1} - x^*\|^2},$$

or equivalently,

$$W_k^{-1} - W_{k-1}^{-1} \geqq \frac{2\lambda_k}{3\|x_{k-1} - x^*\|^2}. \tag{3.3}$$

This is the desired recursive inequality. Summing (3.3) for $k = 1, \cdots, n$, we obtain

$$W_n^{-1} \geqq W_n^{-1} - W_0^{-1} \geqq \frac{2}{3}\sum_{k=1}^{n} \frac{\lambda_k}{\|x_{k-1} - x^*\|^2},$$

which implies

$$f(x_n) - f(x^*) \equiv W_n \leqq \frac{3}{2}\frac{1}{\sum_{k=1}^{n} \lambda_k\|x_{k-1} - x^*\|^{-2}}. \tag{3.4}$$

Multiplying (3.4) by σ_n gives

$$\sigma_n(f(x_n) - f(x^*)) \leqq \frac{3}{2}\frac{1}{\sigma_n^{-1}\sum_{k=1}^{n} \lambda_k\|x_{k-1} - x^*\|^{-2}}.$$

Since $\|x_n - x^*\| \to 0$, $\|x_n - x^*\|^{-1} \to \infty$. Therefore, using the Silverman-Toeplitz theorem (see Dunford and Schwartz [6]), $\sigma_n^{-1}\sum_{k=1}^{n} \lambda_k\|x_{k-1} - x^*\|^{-2} \to \infty$ also. Consequently, $f(x_n) - f(x^*) = o(\sigma_n^{-1})$. □

4. A fundamental estimate. Let $A: H \to H$ be a maximal monotone operator. Consider two proximal trajectories $\{x_k\}_{k=0}^{\infty}$, and $\{\hat{x}_k\}_{k=0}^{\infty}$. A remarkable estimate due to Kobayasi, Kobayashi, and Oharu [8], [9] (see also Pavel [14]) gives an estimate for the distance $\|x_k - \hat{x}_l\|$ between an arbitrary point x_k in the first trajectory and the point $\hat{x}_l$ on the second trajectory. This estimate can be used as the basis for the theory of nonlinear contractive semigroups and nonlinear evolution equations in Banach spaces. We shall use it in §5 to help settle a question posed by Rockafellar [17] on the strong convergence of the PPA. Since Kobayashi's estimate does not seem to be known in the optimization literature, but is likely to have further applications in optimization, we develop a special version of it here which will be enough for our purposes. The interested reader should consult Kobayasi, Kobayashi, and Oharu [9], or Pavel [14] for the general version of the estimate.

We will use two simple lemmas, valid for any monotone operator (not necessarily maximal).

LEMMA 4.1. *If $A: H \to H$ is a monotone operator, then for any $\lambda > 0$, and $y_i \in A(x_i)$, $i = 1,2$,*

$$\|x_1 - x_2\| \leqq \|x_1 - x_2 + \lambda(y_1 - y_2)\|.$$

Lemma 4.1 follows from the application of (1.3) to the operator λA.

LEMMA 4.2. *If A is a monotone operator, then for any $\lambda, \mu \geqq 0$, and $y_i \in A(x_i)$, $i=1,2$, the following inequality holds*:

$$(\lambda+\mu)\|x_1-x_2\| \leqq \lambda\|x_2+\mu y_2-x_1\|+\mu\|x_1+\lambda y_1-x_2\|.$$

Proof. We have

$$\mu(x_1+\lambda y_1-x_2)-\lambda(x_2+\mu y_2-x_1)=(\lambda+\mu)(x_1-x_2)+\lambda\mu(y_1-y_2). \tag{4.1}$$

By Lemma 4.1,

$$\|x_1-x_2\| \leqq \left\|(x_1-x_2)+\frac{\lambda\mu}{\lambda+\mu}(y_1-y_2)\right\|,$$

so that

$$\begin{aligned}(\lambda+\mu)\|x_1-x_2\| &\leqq \|(\lambda+\mu)(x_1-x_2)+\lambda\mu(y_1-y_2)\|\\ &=\|\mu(x_1+\lambda y_1-x_2)-\lambda(x_2+\mu y_2-x_1)\|\\ &\leqq \lambda\|x_2+\mu y_2-x_1\|+\mu\|x_1+\lambda y_1-x_2\|,\end{aligned}$$

where the equality follows from (4.1). The proof is complete. □

We now return to the PPA. In the remainder of this section we assume that a maximal monotone operator $A: H \to H$ is given, and that it generates the proximal trajectory $\{x_k\}_{k=0}^{\infty}$, with $x_0=x$ and parameters $\{\lambda_k\}_{k=1}^{\infty}$.

LEMMA 4.3. *For any $u \in D(A)$, and $k \geqq 0$,*

$$\|x_k-u\| \leqq \|x_0-u\|+\sigma_k\|A^0 u\|. \tag{4.2}$$

Proof. Let v be an arbitrary element of $A(u)$. Since $y_j=(x_{j-1}-x_j)/\lambda_j \in A(x_j)$, we have $x_{j-1}=x_j+\lambda_j y_j$. From Lemma 4.1,

$$\begin{aligned}\|x_j-u\| &\leqq \|x_j-u+\lambda_j(y_j-v)\|=\|x_{j-1}-u-\lambda_j v\|\\ &\leqq \|x_{j-1}-u\|+\lambda_j\|v\|.\end{aligned} \tag{4.3}$$

The lemma follows by summing (4.3) for $j=1,\cdots,k$ and noting that $v \in A(u)$ is arbitrary. □

Consider two proximal trajectories $\{x_k\}_{k=0}^{\infty}$, and $\{\hat{x}_l\}_{l=0}^{\infty}$. We will derive an estimate for the distance $\|x_k-\hat{x}_l\|$ for arbitrary k and l. This estimate will be obtained recursively.

We first prove some preliminary results. Denote the mesh of the first proximal trajectory by $d=\max_{1\leqq k\leqq N}\lambda_k$. Similarly, the mesh of the second proximal trajectory is defined by $\hat{d}=\max_{1\leqq l\leqq \hat{N}}\hat{\lambda}_l$. Also, we define $\alpha_{k,l}=\hat{\lambda}_l/(\lambda_k+\hat{\lambda}_l)$ and $\beta_{k,l}=1-\alpha_{k,l}=\lambda_k/(\lambda_k+\hat{\lambda}_l)$. Finally, we define

$$c_{k,l}=\sqrt{(\sigma_k-\hat{\sigma}_l)^2+d\sigma_k+\hat{d}\hat{\sigma}_l}.$$

LEMMA 4.4. $\alpha_{k,l}c_{k-1,l}+\beta_{k,l}c_{k,l-1} \leqq c_{k,l}$.

Proof. We have

$$\begin{aligned}\alpha_{k,l}c_{k-1,l}+\beta_{k,l}c_{k,l-1} &= \alpha_{k,l}^{1/2}(\alpha_{k,l}^{1/2}c_{k-1,l})+\beta_{k,l}^{1/2}(\beta_{k,l}^{1/2}c_{k,l-1})\\ &\leqq \sqrt{\alpha_{k,l}c_{k-1,l}^2+\beta_{k,l}c_{k,l-1}^2},\end{aligned}$$

where the inequality follows from the Cauchy-Schwarz inequality and the fact that

$\alpha_{k,l}+\beta_{k,l}=1$. Therefore,

$$\begin{aligned}(\alpha_{k,l}c_{k-1,l}+\beta_{k,l}c_{k,l-1})^2 &\leqq \alpha_{k,l}c_{k-1,l}^2+\beta_{k,l}c_{k,l-1}^2\\ &= \alpha_{k,l}((\sigma_{k-1}-\hat{\sigma}_l)^2+d\sigma_{k-1}+\hat{d}\hat{\sigma}_l)\\ &\qquad +\beta_{k,l}((\sigma_k-\hat{\sigma}_{l-1})^2+d\sigma_k+\hat{d}\hat{\sigma}_{l-1})\\ &= \alpha_{k,l}((\sigma_k-\hat{\sigma}_l-\lambda_k)^2+d(\sigma_k-\lambda_k)+\hat{d}\hat{\sigma}_l)\\ &\qquad +\beta_{k,l}((\sigma_k-\hat{\sigma}_l+\hat{\lambda}_l)^2+d\sigma_k+\hat{d}(\hat{\sigma}_l-\hat{\lambda}_l))\\ &= \frac{\hat{\lambda}_l}{\lambda_k+\hat{\lambda}_l}(c_{k,l}^2+\lambda_k^2-2\lambda_k(\sigma_k-\hat{\sigma}_l)-d\lambda_k)\\ &\qquad +\frac{\lambda_k}{\lambda_k+\hat{\lambda}_l}(c_{k,l}^2+\hat{\lambda}_l^2+2\hat{\lambda}_l(\sigma_k-\hat{\sigma}_l)-\hat{d}\hat{\lambda}_l)\\ &= c_{k,l}^2+\frac{\lambda_k\hat{\lambda}_l}{\lambda_k+\hat{\lambda}_l}((\lambda_k-d)+(\hat{\lambda}_l-\hat{d}))\leqq c_{k,l}^2.\end{aligned}$$

The lemma is proved. □

THEOREM (Kobayashi) 4.1. *Let $u\in D(A)$ be an arbitrary point. Then, for any $k=0,\cdots,N$ and $l=0,\cdots,\hat{N}$,*

$$\|x_k-\hat{x}_l\|\leqq\|x_0-u\|+\|\hat{x}_0-u\|+\sqrt{(\sigma_k-\hat{\sigma}_l)^2+d\sigma_k+\hat{d}\hat{\sigma}_l}\cdot\|A^0u\|.$$

Proof. Observe that the coefficient of $\|A^0u\|$ in the desired estimate is simply $c_{k,l}$. The proof will be by induction. We start by proving the theorem for the pairs $(k,0)$ and $(0,l)$. We have

$$\begin{aligned}\|x_k-\hat{x}_0\| &\leqq \|x_k-u\|+\|\hat{x}_0-u\|\\ &\leqq \|x_0-u\|+\|\hat{x}_0-u\|+\sigma_k\|A^0u\|\\ &\leqq \|x_0-u\|+\|\hat{x}_0-u\|+c_{k,0}\|A^0u\|,\end{aligned}$$

where the second inequality follows from Lemma 4.3, and the last inequality follows from the fact that $\sigma_k\leqq c_{k,0}$, which is easy to see. This proves the theorem for $(k,0)$. By symmetry, the theorem is also true for the pair $(0,l)$.

Suppose we have proved the theorem for the pairs $(k-1,l)$ and $(k,l-1)$. From Lemma 4.2, we obtain

$$(\lambda_k+\hat{\lambda}_l)\|x_k-\hat{x}_l\|\leqq\lambda_k\|\hat{x}_l+\hat{\lambda}_l\hat{y}_l-x_k\|+\hat{\lambda}_l\|x_k+\lambda_k y_k-\hat{x}_l\|.$$

Noting that $\hat{x}_{l-1}=\hat{x}_l+\hat{\lambda}_l\hat{y}_l$ and $x_{k-1}=x_k+\lambda_k y_k$, we have

$$\begin{aligned}\|x_k-\hat{x}_t\| &\leqq \frac{\hat{\lambda}_l}{\lambda_k+\hat{\lambda}_l}\|x_{k-1}-\hat{x}_l\|+\frac{\lambda_k}{\lambda_k+\hat{\lambda}_l}\|x_k-\hat{x}_{l-1}\|\\ &= \alpha_{k,l}\|x_{k-1}-\hat{x}_l\|+\beta_{k,l}\|x_k-\hat{x}_{l-1}\|\\ &\leqq \alpha_{k,l}(\|x_0-u\|+\|\hat{x}_0-u\|+c_{k-1,l}\|A^0u\|)\\ &\qquad +\beta_{k,l}(\|x_0-u\|+\|\hat{x}_0-u\|+c_{k,l-1}\|A^0u\|)\\ &\leqq \|x_0-u\|+\|\hat{x}_0-u\|+(\alpha_{k,l}c_{k-1,l}+\beta_{k,l}c_{k,l-1})\|A^0u\|\\ &\leqq \|x_0-u\|+\|\hat{x}_0-u\|+c_{k,l}\|A^0u\|,\end{aligned}$$

where the second inequality follows from the induction hypothesis, and the last inequality follows from Lemma 4.4. The theorem is proved. □

Theorem 4.1 can be used to prove that as the mesh of the backward discretization $\max_{k\geqq 1} \lambda_k \to 0$ in (1.6), the proximal trajectory converges to a (unique) *discrete scheme* solution to the differential inclusion (1.5). It turns out that this solution coincides with the usual solution $u(t)=S(t)x$ (also called the strong solution) discussed in §1. See Kobayashi [8], or Pavel [14, Chap. 1, §3] for more details.

COROLLARY 4.1. *The following estimates hold*:

$$\|x_k - S(t)x\| \leqq \sqrt{(\sigma_k - t)^2 + d\sigma_k} \cdot \|A^0x\|, \tag{4.4}$$

$$\|S(t)x - S(s)x\| \leqq |t-s| \cdot \|A^0x\|. \tag{4.5}$$

Proof. Choose $\hat{x}_0 = x$ and apply Theorem 4.1. As $\hat{d}\to 0$, the second proximal trajectory converges to the continuous path $S(t)x$, and (4.4) follows. Estimate (4.5) is proved similarly. □

5. On the strong convergence of the proximal point algorithm. We noted in §2 that the trajectory of the PPA converges weakly to a minimizer of a proper, closed convex function, provided that $X^* \neq \varnothing$ and $\sigma_k \to \infty$. In [17], Rockafellar posed the question of whether weak convergence can be strengthened to strong convergence. This question is also important for us since strong convergence has a bearing on the rate of convergence of the PPA. By Theorem 2.1, $f(x_n) - f^* = O(\sigma_n^{-1})$ in the case of weak convergence; however, by Theorem 3.1 $f(x_n) - f^* = o(\sigma_n^{-1})$ in the case of strong convergence. Of course, in finite dimensions, weak and strong convergence are equivalent. There are also cases (see, for example, [4], [5]) where we can show strong convergence.

In this section, we answer Rockafellar's open question in the negative. In particular, we prove that in l^2 there is a function f such that given any positive bounded sequence $\{\lambda_j\}_{j=1}^\infty$, there is a starting point $x \in D(f)$, and the PPA starting from x with $x_{k+1} = (x_k)_{\lambda_{k+1}}$ converges weakly, but not strongly.

We proceed in the following way. A well-known result of Bruck [5] states that $S(t)x$ converges weakly to a minimizer of f. Baillon [2], following a suggestion of Komura, constructed a proper, closed convex function in l^2 and a point $x \in \overline{D(f)}$ such that $S(t)x$ converges weakly, but not strongly, to a minimizer of f. In [13], Passty showed that the strong (respectively, weak) convergence of $S(t)x$ is equivalent to the strong (respectively, weak) convergence of a PPA trajectory under very restrictive conditions. By using the special properties of the monotone operator ∂f outlined in §2, and the fundamental estimate of Kobayashi described in §4, we show the asymptotic equivalence of the trajectory of a PPA and $S(t)$ under the condition that the sequence $\{\lambda_k\}_{k=1}^\infty$ is bounded.

DEFINITION 5.1. Let C be convex subset of H. A contractive evolution system on C is a two-parameter family of maps $\{U(t,s)\colon 0 \leqq s \leqq t\}$ from C into C satisfying:

(i) $U(t,t)x = x$ for all $x \in C$ and $t \geqq 0$,

(ii) $U(t,s)U(s,r)x = U(t,r)x$ for all $x \in C$ and $0 \leqq r \leqq s \leqq t$, and

(iii) $\|U(t,s)x - U(t,s)y\| \leqq \|x-y\|$ for all $x, y \in C$ and $0 \leqq s \leqq t$.

DEFINITION 5.2. A contractive evolution system $U(t,s)$ is asymptotically equal to a contractive semigroup $S(t)$ if, for all $x \in C$, we have

(i) $\lim_{t\to\infty} \|U(t+h,s)x - S(h)U(t,s)x\| = 0$ for all $t \geqq 0$, uniformly in $h \geqq 0$, and

(ii) $\lim_{t\to\infty} \|U(t+h,t)S(t)x - S(t+h)x\| = 0$ uniformly in $h \geqq 0$. The system U is called an asymptotic semigroup if there is a semigroup to which it is asymptotically equal.

Intuitively, that U and S are asymptotically equal means the following: if we follow one of the trajectories, say S, for a sufficiently long time t and arrive at the point $S(t)x$, then it matters little whether we follow S or U for any length of time h in the future, because the two trajectories will be close to each other.

The concept of asymptotic equality is important because of the following result proved in Passty [13].

LEMMA 5.1. *Let A be a maximal monotone operator on H, and let $S(t)$ be the contractive semigroup generated by A on $\bar{C}$. Let $U(t, s)$ be a contractive evolution system which is asymptotically equal to $S(t)$ on $D(A)$. Then the following are equivalent*:

(i) *$S(t)x$ converges strongly* (*respectively weakly*) *as $t\to\infty$ for all $x\in D(A)$*,

(ii) *$U(t, s)x$ converges strongly* (*respectively weakly*) *as $t\to\infty$ for all $x\in D(A)$, and $s\geqq 0$.*

Remark 5.1. From the proof given in Passty [13], it can be seen that in proving (i) implies (ii) in the lemma, only condition (i) of Definition 5.2 is needed. Similarly, only condition (ii) of Definition 5.2 is needed for proving that (ii) implies (i).

LEMMA 5.2. *Let $\{k(n)\}_{n=0}^{\infty}$ be a sequence of strictly increasing positive integers, where $k(0)=0$. Define $\sigma_m^n=\sum_{j=k(m)+1}^{k(n)}\lambda_j$, and $\prod_m^n=\prod_{j=k(m)+1}^{k(n)}J_{\lambda_j}$.* (*If $n\leqq m$, we define $\sigma_m^n=0$ and $\prod_m^n x=x$.*) *Then, for any $n, p\geqq 1$,*

$$(5.1)\qquad \left\|S(\sigma_n^{n+p})x-\prod_n^{n+p}x\right\|\leqq\sum_{m=n+1}^{n+p}\left\|S(\sigma_{m-1}^m)\left(\prod_n^{m-1}x\right)-\prod_{m-1}^m\left(\prod_n^{m-1}x\right)\right\|,$$

$$(5.2)\qquad \left\|S(\sigma_n^{n+p})x-\prod_n^{n+p}x\right\|\leqq\sum_{m=n+1}^{n+p}\left\|S(\sigma_{m-1}^m)S(\sigma_n^{m-1})x-\prod_{m-1}^m S(\sigma_n^{m-1})x\right\|.$$

Proof. We prove the lemma by induction on p. We first prove (5.1). It is evidently true for $p=1$. Assuming it is true for p, we prove it for $p+1$. We have

$$\begin{aligned}
\left\|S(\sigma_n^{n+p+1})x-\prod_n^{n+p+1}x\right\| &= \left\|S(\sigma_{n+p}^{n+p+1})S(\sigma_n^{n+p})x-\prod_{n+p}^{n+p+1}\left(\prod_n^{n+p}x\right)\right\| \\
&\leqq \left\|S(\sigma_{n+p}^{n+p+1})S(\sigma_n^{n+p})x-S(\sigma_{n+p}^{n+p+1})\left(\prod_n^{n+p}x\right)\right\| \\
&\quad+\left\|S(\sigma_{n+p}^{n+p+1})\left(\prod_n^{n+p}x\right)-\prod_{n+p}^{n+p+1}\left(\prod_n^{n+p}x\right)\right\| \\
&\leqq \left\|S(\sigma_n^{n+p})x-\prod_n^{n+p}x\right\|+\left\|S(\sigma_{n+p}^{n+p+1})\left(\prod_n^{n+p}x\right)-\prod_{n+p}^{n+p+1}\left(\prod_n^{n+p}x\right)\right\| \\
&\leqq \sum_{m=n+1}^{n+p}\left\|S(\sigma_{m-1}^m)\left(\prod_n^{m-1}x\right)-\prod_{m-1}^m\left(\prod_n^{m-1}x\right)\right\| \\
&\quad+\left\|S(\sigma_{n+p}^{n+p+1})\left(\prod_n^{n+p}x\right)-\prod_{n+p}^{n+p+1}\left(\prod_n^{n+p}x\right)\right\| \\
&= \sum_{m=n+1}^{n+p+1}\left\|S(\sigma_{m-1}^m)\left(\prod_n^{m-1}x\right)-\prod_{m-1}^m\left(\prod_n^{m-1}x\right)\right\|,
\end{aligned}$$

where the second inequality follows since $S(t)$ is contractive, and the last inequality follows from the induction hypothesis. This proves (5.1).

We now prove (5.2). The proof is again by induction on p. Equation (5.2) is clearly true for $p=1$. Assuming it is true for p, we prove it for $p+1$. We have

$$\begin{aligned}
\left\| S(\sigma_n^{n+p+1})x-\prod_n^{n+p+1} x\right\| &= \left\| \prod_{n+p}^{n+p+1}\prod_n^{n+p} x-S(\sigma_{n+p}^{n+p+1})S(\sigma_n^{n+p})x\right\| \\
&\leqq \left\| \prod_{n+p}^{n+p+1}\left(\prod_n^{n+p} x\right)-\prod_{n+p}^{n+p+1} S(\sigma_n^{n+p})x\right\| \\
&\quad + \left\| \prod_{n+p}^{n+p+1} S(\sigma_n^{n+p})x-S(\sigma_{n+p}^{n+p+1})S(\sigma_n^{n+p})x\right\| \\
&\leqq \left\| \prod_n^{n+p} x-S(\sigma_n^{n+p})x\right\| + \left\| \prod_{n+p}^{n+p+1} S(\sigma_n^{n+p})x \right. \\
&\qquad\qquad \left. -S(\sigma_{n+p}^{n+p+1})S(\sigma_n^{n+p})x\right\| \\
&\leqq \sum_{m=n+1}^{n+p} \left\| S(\sigma_{m-1}^m)S(\sigma_n^{m-1})x-\prod_{m-1}^m S(\sigma_n^{m-1})x\right\| \\
&\quad + \left\| \prod_{n+p}^{n+p+1} S(\sigma_n^{n+p})x-S(\sigma_{n+p}^{n+p+1})S(\sigma_n^{n+p})x\right\| \\
&= \sum_{m=n+1}^{n+p+1} \left\| S(\sigma_{m-1}^m)S(\sigma_n^{m-1})x-\prod_{m-1}^m S(\sigma_n^{m-1})x\right\|,
\end{aligned}$$

where the second inequality follows since $\prod_{n+p}^{n+p+1}$ is contractive, and the last inequality follows from the induction hypothesis. This proves (5.2). □

Let the sequence $\{\lambda_j\}_{j=1}^{\infty}$ of positive numbers be the parameters of a PPA such that $\sigma_n\to\infty$. We *define* an integer-valued function $n(t)$ for $t\geqq 0$ as follows: $n(0)=0$, and for $t>0$, $n(t)$ is the integer satisfying

$$\sigma_{n(t)-1}<t\leqq\sigma_{n(t)}.$$

We are interested in the contractive evolution system $U(t,s)$, defined for $0\leqq s\leqq t$ by the formula $U(t,s)x=(\prod_{i=n(s)+1}^{n(t)} J_{\lambda_i})x$, where we let $(\prod_{i=1}^{0} J_{\lambda_i})x=x$. It is easy to show that $U(t,s)$ is a contractive evolution system using the fact that J_λ is a contractive mapping.

The following theorem is the main result of this section. This sharpens Theorem 1 in Passty [13] in the case where $A=\partial f$ in that our conditions on the parameters $\{\lambda_k\}_{k=1}^{\infty}$ are much more relaxed than Passty's. Our relaxed conditions are possible because of the special properties of the operator $A=\partial f$ given in §2. However, Passty's Theorem 1 applies to an arbitrary maximal monotone operator.

THEOREM 5.1. *Suppose* $f: H\to\mathbf{R}\cup\{\infty\}$ *is a closed convex function and assume that* f *has a minimizer. Let* $\{\lambda_j\}_{j=1}^{\infty}$ *be a bounded sequence of positive numbers such that* $\sigma_n\to\infty$. *Then the contractive evolution system* $U(t,s)$ *defined above is asymptotically equal to the contractive semigroup* $S(t)$ *generated by* ∂f.

Proof. By Lemma 5.1, we need to verify the conditions (i) and (ii) in Definition 5.2.

For the sake of simple notation, we define $\sigma_s^t = \sum_{j=n(s)+1}^{n(t)} \lambda_j$, and $\prod_s^t = \prod_{j=n(s)+1}^{n(t)} J_{\lambda_j}$. If $t \leqq s$, we let $\sigma_s^t = 0$, and $\prod_s^t x = x$. Note that we also define σ_m^n in Lemma 5.2. However, no confusion should arise, since it will be clear from the context which definition is intended.

Let us first verify condition (i) in Definition 5.2. Without loss of generality, we may assume that $s = 0$. Fix $t > 0$. For an arbitrary $h > 0$, we have

$$\begin{aligned}
\left\| U(t+h,0)x - S(h)U(t,0)x \right\| &= \left\| \prod_0^{t+h} x - S(h)\prod_0^t x \right\| \\
&= \left\| \prod_t^{t+h}\left(\prod_0^t x\right) - S(h)\left(\prod_0^t x\right) \right\| \\
&\leqq \left\| \prod_t^{t+h}\left(\prod_0^t x\right) - S(\sigma_t^{t+h})\left(\prod_0^t x\right) \right\| \\
&\quad + \left\| S(\sigma_t^{t+h})\left(\prod_0^t x\right) - S(h)\prod_0^t x \right\|.
\end{aligned}$$

We first estimate the second term in the last expression above:

$$\begin{aligned}
\left\| S(\sigma_t^{t+h})\prod_0^t x - S(h)\prod_0^t x \right\| &\leqq |\sigma_t^{t+h} - h| \cdot \left\| A^0 \prod_0^t x \right\| \\
&\leqq \frac{|\sigma_t^{t+h} - h|}{\sigma_{n(t)}} \rho(x, X^*) \\
&\leqq \frac{\max\{\lambda_{n(t)}, \lambda_{n(t+h)}\}}{\sigma_{n(t)}} \rho(x, X^*) \qquad (5.3)\\
&\leqq \frac{\Lambda}{\sigma_{n(t)}} \rho(x, X^*),
\end{aligned}$$

where $\Lambda = \max_{j \leqq 1} \lambda_j$. Here the first inequality follows from (4.5), and the second inequality follows from (2.12) with x^* replacing u, where x^* is the element of X^* closest to x. The third inequality follows easily from the definition of σ_t^{t+h}. Since $\sigma_{n(t)} \to \infty$ as $t \to \infty$, the last term in (5.3) can be made as small as desired by choosing t large enough.

It remains to estimate the first term:

$$\left\| S(\sigma_t^{t+h})\left(\prod_0^t x\right) - \prod_t^{t+h}\left(\prod_0^t x\right) \right\|.$$

The idea is to *partition* the interval $[0, \sigma_{n(t+h)}]$ into subintervals and use Lemma 5.2 on each subinterval. The subintervals will be of the form $[\sigma_{k(i)}, \sigma_{k(i+1)}]$, for $i = 0, \cdots, n+p$, such that $k(i) = n(t)$ for some t, where we assume $n(t) = k(n)$ (note the two meanings of n here) and $n(t+h) = k(n+p)$. We will impose more conditions on the sequence $k(i)$ later. We have

$$
\begin{aligned}
\left\| S(\sigma_t^{t+h})\left(\prod_0^t x\right) - \prod_t^{t+h}\left(\prod_0^t x\right)\right\| &= \left\| S(\sigma_n^{n+p})\left(\prod_0^n x\right) - \prod_n^{n+p}\left(\prod_0^n x\right)\right\| \\
&\leqq \sum_{m=n+1}^{n+p} \left\| S(\sigma_{m-1}^m) \prod_n^{m-1}\left(\prod_0^n x\right) - \prod_{m-1}^{m}\prod_n^{m-1}\left(\prod_0^n x\right)\right\| \\
&= \sum_{m=n+1}^{n+p} \left\| S(\sigma_{m-1}^m)\left(\prod_0^{m-1} x\right) - \prod_{m-1}^{m}\left(\prod_0^{m-1} x\right)\right\| \\
&\leqq \sum_{m=n+1}^{n+p} \sqrt{d_m \sigma_{m-1}^m}\left\| A^0\left(\prod_0^{m-1} x\right)\right\| \\
&\leqq \sum_{m=n+1}^{n+p} \sqrt{d_m \sigma_{m-1}^m}\,\frac{\rho(x, X^*)}{\sigma_0^{m-1}} \\
&\leqq \sqrt{\Lambda}\rho(x, X^*) \sum_{m=n+1}^{n+p} \frac{\sqrt{\sigma_{m-1}^m}}{\sigma_0^{m-1}},
\end{aligned}
$$

where the first inequality follows from (5.1), and the second inequality from (4.4). The third inequality follows from (2.12), with x^* replacing u, where x^* is the element of X^* closest to x. Here $d_m = \max_{j=k(m-1)+1}^{k(m)} \lambda_j$. We are interested in making the term

$$
\sum_{m=n+1}^{n+p} \frac{\sqrt{\sigma_{m-1}^m}}{\sigma_0^{m-1}}
$$

small. Clearly, if $t \to \infty$, then $n(t) \to \infty$ also. Therefore, if we can ensure that

$$
\sum_{m=1}^{\infty} \frac{\sqrt{\sigma_{m-1}^m}}{\sigma_0^{m-1}} < \infty,
$$

we are done. There are many choices for $k(i)$ which can accomplish this. For example, if we choose $k(i)$ such that

$$
\sigma_{k(i)-1} < i^2 \leqq \sigma_{k(i)} \tag{5.4}
$$

we can easily check that

$$
\sqrt{\sigma_{m-1}^m}/\sigma_0^{m-1} \leqq \sqrt{2m-1+\Lambda}/(m-1)^2
$$

and therefore the infinite series above converges. This proves (i) of Definition 5.2.

Next, we need to verify condition (ii) in Definition 5.2. We have

$$
\begin{aligned}
\|U(t+h, t)S(t)x - S(t+h)x\| &= \left\| \prod_t^{t+h} S(t)x - S(t+h)x \right\| \\
&\leqq \left\| \prod_t^{t+h} S(t)x - S(\sigma_t^{t+h})S(t)x \right\| \\
&\quad + \|S(\sigma_t^{t+h})S(t)x - S(h)S(t)x\|.
\end{aligned}
$$

The second term above can be estimated as follows:

$$
\begin{aligned}
\|S(\sigma_t^{t+h})x - S(h)S(t)x\| &\leqq |\sigma_t^{t+h} - h| \cdot \|A^0 S(t)x\| \\
&\leqq \frac{|\sigma_t^{t+h} - h|}{t}\,\rho(x, X^*) \qquad (5.5) \\
&\leqq \frac{\Lambda}{t}\,\rho(x, X^*),
\end{aligned}
$$

where the first inequality follows from (4.5), and the second one from (2.14). Therefore, the last term in (5.5) can be made as small as desired by choosing t large enough.

Finally, we estimate the remaining term:

$$\begin{aligned}\left\| \prod_{t}^{t+h} S(t)x - S(\sigma_t^{t+h})S(t)x \right\| &= \left\| \prod_{n}^{n+p} S(t)x - S(\sigma_n^{n+p})S(t)x \right\| \\ &\leqq \sum_{m=n+1}^{n+p} \left\| S(\sigma_{m-1}^m)S(\sigma_n^{m-1})S(t)x - \prod_{m-1}^{m} S(\sigma_n^{m-1})S(t)x \right\| \\ &= \sum_{m=n+1}^{n+p} \left\| S(\sigma_{m-1}^m)S(\sigma_n^{m-1}+t)x - \prod_{m-1}^{m} S(\sigma_n^{m-1}+t)x \right\| \\ &\leqq \sum_{m=n+1}^{n+p} \sqrt{d_m\sigma_{m-1}^m}\|A^0S(\sigma_n^{m-1}+t)x\| \\ &\leqq \sum_{m=n+1}^{n+p} \sqrt{d_m\sigma_{m-1}^m}\frac{\rho(x, X^*)}{\sigma_n^{m-1}+t} \\ &\leqq \sqrt{\Lambda}\rho(x, X^*) \sum_{m=n+1}^{n+p} \frac{\sqrt{\sigma_{m-1}^m}}{\sigma_0^{m-1}-\lambda_{k(n)}} \\ &\leqq \sqrt{\Lambda}\rho(x, X^*) \sum_{m=n+1}^{n+p} \frac{\sqrt{\sigma_{m-1}^m}}{\sigma_0^{m-1}-\Lambda},\end{aligned}$$

where the first inequality follows from (5.2), the second inequality from (4.4), and the third one from (2.14). If $k(i)$ is chosen as in (5.4), it is easy to check, as above, that

$$\sum_{m=n+1}^{n+p} \frac{\sqrt{\sigma_{m-1}^m}}{\sigma_0^{m-1}-\Lambda} \leqq \sum_{m=n+1}^{n+p} \frac{\sqrt{2m-1+\Lambda}}{(m-1)^2-\Lambda} \to 0$$

as $n \to \infty$. This proves (ii) of Definition 5.2. □

COROLLARY 5.1. *There exists a proper, closed convex function f in l^2 such that given any bounded positive sequence $\{\lambda_j\}_{j=1}^{\infty}$, there exists a point $x \in D(f)$ for which* PPA *starting at x, $x_{k+1} = (x_k)_{\lambda_{k+1}}$ converges weakly, but not strongly to a minimizing point of f.*

Proof. By Baillon's theorem [2], there exists a function f in $H = l^2$ and a starting point x such that $S(t)x$ converges weakly but not strongly to a minimizer of f. By Theorem 5.1, $U(t, s)$, defined above, is asymptotically equivalent to $S(t)$. Therefore, by Lemma 5.1, there exists a point $\bar{x}$ such that $U(t, s)\bar{x}$ also converges weakly but not strongly to a minimizer of f. □

Acknowledgments. The author thanks the two anonymous referees for their helpful comments.

REFERENCES

[1] J. P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, Interscience Publications, John Wiley, New York, 1984.

[2] J. B. BAILLON, *Un exemple concernant le comportement asymptotique de la solution due problème $du/dt + \partial\varphi(u) \ni 0$*, J. Funct. Anal., 28 (1978), pp. 369-376.

[3] H. BRÈZIS, *Opérateurs Maximaux Monotones*, Mathematics Studies No. 5, North-Holland, Amsterdam 1973.

[4] H. BRÉZIS AND P. L. LIONS, *Produits infinis de résolvantes*, Israel J. Math., 29 (1978), pp. 329-345.

[5] R. E. Bruck, *Asymptotic convergence of nonlinear contraction semigroups in Hilbert spaces*, J. Funct. Anal., 18 (1975), pp. 15-26.

[6] N. Dunford and J. T. Schwartz, *Linear Operators, Part* I: *General Theory*, Interscience Publications, John Wiley, New York, 1988.

[7] O. Güler, *New proximal point algorithms for convex minimization*, Math. Programming, submitted.

[8] Y. Kobayashi, *Difference approximation of Cauchy problems for quasidissipative operators and generation of nonlinear semigroups*, J. Math. Society Japan, 27 (1975), 640-665.

[9] K. Kobayasi, Y. Kobayashi, and S. Oharu, *Nonlinear evolution operators in Banach spaces*, Osaka. J. Math., 21 (1984), pp. 281-310.

[10] B. Martinet, *Regularisation, d'inéquations variationelles par approximations succesives*, Revue Française d'Informatique et de Recherche Operationelle, 1970, pp. 154-159.

[11] G. Minty, *Monotone (nonlinear) operators in a Hilbert space*, Duke Math. J., 29 (1962), pp. 341-348.

[12] J. J. Moreau, *Proximité et dualité dans un espace Hilbertien*, Bull. Soc. Math., France, 93 (1965), pp. 273-299.

[13] G. B. Passty, *Preservation of the asymptotic behavior of a nonlinear contraction semigroup by backward differencing*, Houston J. Math., 7 (1981), pp. 103-110.

[14] N. H. Pavel, *Nonlinear Evolution Operators and Semigroups*, Lecture Notes in Mathematics, Vol. 1260, Springer-Verlag, New York, 1987.

[15] S. Reich, *On infinite products of resolvents*, Atti. Accad. Naz. Lincei, 63 (1977), pp. 338-340.

[16] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[17] ———, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877-898.

[18] ———, *Augmented Lagrangians and applications of the proximal point algorithm in convex programming*, Math. Oper. Res., 1 (1976), pp. 97-116.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 420-444, March 1991

REALIZATION OF ACAUSAL WEIGHTING PATTERNS WITH BOUNDARY-VALUE DESCRIPTOR SYSTEMS*

RAMINE NIKOUKHAH†, BERNARD C. LEVY‡, AND ALAN S. WILLSKY§

Abstract. This paper examines the realization of acausal weighting patterns with two-point boundary-value descriptor systems (TPBVDSs). Attention is restricted to the subclass of TPBVDSs that are *stationary*, so that their input-output weighting pattern is shift-invariant, and *extendible*, i.e., their weighting pattern can be extended outwards indefinitely. Then, given an infinite acausal shift-invariant weighting pattern, the realization problem consists of constructing a minimal TPBVDS over a fixed interval, whose extended weighting pattern matches the given pattern. The realization method that is proposed relies on a new transform, the (s, t)-transform, which is better adapted to the analysis of descriptor dynamics than the standard z-transform, since it handles zero and infinite frequencies in a symmetric way. This new transform is used to determine the dimension of a minimal realization, and then to construct a minimal realization by obtaining state-space representations for two homogeneous rational matrices in s and t obtained from the causal and anticausal components of the weighting pattern.

Key words. acausal weighting pattern, boundary-value descriptor system, realization theory, (s, t) transform, McMillan degree

AMS(MOS) subject classifications. 93B15, 93B20

1. Introduction.

There exists an extensive literature [1]-[4] on the state-space realization problem for linear time-invariant causal systems, i.e., for systems which admit an input-output description of the form

$$y(k)=\sum_{l=-\infty}^{\infty} W(k-l)u(l), \tag{1.1}$$

where the impulse response (weighting pattern) $W(.)$ satisfies

$$W(k)=0 \quad \text{for } k\leqq 0. \tag{1.2}$$

However, for many physical systems, in particular when the independent variable is space rather than time, the causality condition (1.2) does not hold. For example, if we consider the temperature of a heated rod, there is no reason to assume that the temperature at any point of the rod depends exclusively on the applied heat on one side of that point. Weighting patterns that do not satisfy (1.2) are called acausal. The objective of this paper is to develop a realization theory for acausal weighting patterns in terms of *two-point boundary-value descriptor systems* (TPBVDSs) of the form

$$Ex(k+1)=Ax(k)+Bu(k), \qquad 0\leqq k\leqq N-1, \tag{1.3}$$

with boundary condition

$$V_i x(0)+V_f x(N)=v, \tag{1.4}$$

* Received by the editors August 28, 1989; accepted for publication (in revised form) April 3, 1990. This was supported by the Institut National de Recherche en Informatique et Automatique, Air Force Office of Scientific Research grant AFOSR-88-0032, and National Science Foundation grant ECS-8700903.

† Institut National de Recherche en Informatique et Automatique (INRIA), Rocquencourt, B.P. 105, 78153 Le Chesnay Cedex, France.

‡ Department of Electrical Engineering and Computer Science, University of California, Davis, California 95616.

§ Department of Electrical Engineering and Computer Science and Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

and output

$$y(k) = Cx(k), \qquad 0 \leqq k \leqq N. \tag{1.5}$$

The motivation for considering this class of systems is that the discrete-time descriptor dynamics (1.3) are noncausal, in the sense that they contain components which propagate in both time directions [5]. The boundary conditions (1.4) are another source of noncausality, since they are expressed symmetrically in terms of the system variables at both ends of the interval $[0, N]$. Thus, TPBVDSs have a totally acausal structure which is ideally suited to model noncausal systems [6]-[8]. Motivated by the earlier work of Krener [9]-[10], and Gohberg, Kaashoek, and Lerer [11]-[13] for boundary-value systems with standard nondescriptor dynamics, a complete system theory of TPBVDSs has been developed in [14]-[18], including concepts such as reachability, observability, and minimality. In this paper, we restrict our attention to stationary and extendible TPBVDSs, namely TPBVDSs whose weighting pattern is shift-invariant, and where the interval of definition $[0, N]$ of the TPBVDS can be extended outwards indefinitely, without changing the weighting pattern. This extension process yields an extended weighting pattern $W(k)$ defined for all $k \in \mathbf{Z}$, where the weighting pattern of the original TPBVDS and of all its extensions are restrictions of $W(k)$.

The realization problem that we consider can be stated as follows. Given a weighting pattern $W(k)$, construct a minimal TPBVDS over a sufficiently large interval $[0, N]$, which has $W(k)$ as its extended weighting pattern. As for causal time-invariant systems, where the z-transform plays a useful role in transforming the realization problem into a state-space representation problem for proper rational matrix transfer functions, it is shown that the TPBVDS realization problem can be formulated in the frequency domain as a state-space representation problem for rational transfer functions. However, instead of using the z-transform, we introduce a new transform, the (s, t)-transform, which handles zero and infinite frequencies symmetrically, and is therefore well adapted to the analysis of descriptor systems. Specifically, the (s, t)-transform of a matrix sequence $H(k)$ is defined as

$$H(s, t) = \sum_{k=-\infty}^{\infty} H(k) t^{k-1}/s^k. \tag{1.6}$$

Because of its special structure, $H(s, t)$ is strictly proper when viewed as a function of both s and t, but not necessarily strictly proper in s and t separately. When $H(s, t)$ is rational, this last observation leads us to construct minimal state-space representations of the form

$$H(s, t) = K(sD - tF)^{-1}G, \tag{1.7}$$

where the descriptor dynamics appearing in (1.7) generalize the causal dynamics that are usually employed for strictly proper rational matrices in z.

The (s, t) transform is used here to characterize the dimension of TPBVDS realizations in terms of the McMillan degree of rational matrices in s and t, and to formulate the TPBVDS realization problem as a state-space realization in the (s, t)-domain. More precisely, if $W_f(s, t)$ and $W_b(s, t)$ denote the (s, t)-transforms of the causal and anticausal parts of the weighting pattern $W(k)$, and if

$$W(s, t) = W_f(s, t) + W_b(s, t), \tag{1.8a}$$

$$H_r(s, t) = [\, W_f(s, t)\, W_b(s, t)], \qquad H_o(s, t) = \begin{bmatrix} W_f(s, t) \\ W_b(s, t) \end{bmatrix}, \tag{1.8b}$$

it is shown that minimal TPBVDS realizations of the extended weighting pattern $W(k)$ have dimension

$$n=\omega+\rho-\tau, \tag{1.9}$$

where ω, ρ, and τ denote the McMillan degrees of $H_r(s,t)$, $H_o(s,t)$, and $W(s,t)$, respectively. We also develop a minimal realization procedure, which relies on constructing minimal state-space representations of the form (1.7) for *both* $H_r(s,t)$ and $H_o(s,t)$. The reason why it is necessary to construct state-space representations for two rational matrices, instead of one for the causal case, is that the TPBVDS realization problem requires finding descriptor dynamics (1.3) and boundary conditions (1.4), which together realize $W(k)$. It is the search for boundary conditions that makes the TPBVDS realization problem significantly harder than the causal problem.

This paper is organized as follows. In § 2, we review several results concerning the stationarity, minimality, and extendibility of TPBVDSs that will be used later. It is shown in § 3 that the effect of the boundary conditions on the extended weighting pattern of the system can be characterized completely by a single matrix, called the decomposition matrix, which appears as a parameter of both the causal and anticausal parts of $W(k)$. This matrix simplifies significantly the presentation of our realization results. In § 4, we examine a direct but naive TPBVDS realization procedure consisting in constructing separate minimal realizations of the causal and anticausal components of $W(k)$. Although the resulting realization is generally nonminimal, it is minimal when the weighting pattern $W(k)$ is summable. Furthermore, it yields necessary and sufficient conditions for the realizability of acausal weighting patterns. The (s,t)-transform is introduced in § 5 and is used to formulate the TPBVDS realization problem in the frequency domain. A method for constructing minimal state-space representations of the form (1.7) for rational matrices in s and t is also presented. Finally, § 6 contains the two main results of our paper, namely the characterization (1.9) for the dimension of a minimal realization, and a minimal TPBVDS realization procedure in the frequency domain.

2. Two-point boundary-value descriptor systems. In this section, we review several properties of TPBVDSs, such as stationarity, minimality, and extendibility, that will be needed in the development of our TPBVDS realization procedure.

2.1. Model description. Consider a linear time-invariant TPBVDS of the form (1.3)-(1.5), where x and v are n-dimensional, u is m-dimensional, y is p-dimensional, and E, A, B, and C are constant matrices. We assume that the length N of the interval of definition satisfies $N\geqq 2n$, so that all modes can be excited and observed. In [14] it was shown that if the system (1.3)-(1.4) is well posed, by left multiplication of (1.3) and (1.4) with invertible matrices, we can bring this system to the following *normalized form*, where there exists scalars α and β such that

$$\alpha E+\beta A=I \tag{2.1}$$

(this is referred to as the *standard form* for the pencil $\{E, A\}$), and

$$V_iE^N+V_fA^N=I. \tag{2.2}$$

Note that (2.1) implies that E and A commute, that E, A, and the system have a common set of eigenvectors,[1] and that $\{E^k, A^k\}$ is a regular pencil for all $k\geqq 0$. Another

[1] v is an eigenvector of the system if $v\neq 0$ and for some σ, $(\sigma E-A)v=0$. σ is called an eigenmode of the system; for descriptor systems σ can be ∞.

consequence of (2.1) is that the space of matrices $\{A^K E^L;\ K, L \geqq 0\}$ is spanned by the n matrices $\{A^k E^{n-1-k};\ 0 \leqq k \leqq n-1\}$. This last result, which was derived in [14], is a generalization of the Cayley-Hamilton theorem to matrix pencils in the standard form (2.1). We assume throughout this paper that (2.1) and (2.2) hold.

As derived in [14], the map from $\{\mathbf{u}, v\}$ to $\mathbf{x}$ has the form:

$$x(k) = A^k E^{N-k} v + \sum_{l=0}^{N-1} G(k, l) B u(l), \tag{2.3}$$

where the Green function $G(k, l)$ is given by

$$G(k, l) = \begin{cases} A^k[A - E^{N-k}(V_i A + \omega V_f E) E^k] E^{l-k} A^{N-l-1} \Gamma^{-1}, & l \geqq k, \\ E^{N-k}[\omega E - A^k(V_i A + \omega V_f E) A^{N-k}] E^l A^{k-l-1} \Gamma^{-1}, & l < k, \end{cases} \tag{2.4}$$

and where ω is any number such that

$$\Gamma = \omega E^{N+1} - A^{N+1} \tag{2.5}$$

is invertible.

The map from inputs $\mathbf{u}$ to outputs $\mathbf{y}$ specifies the weighting pattern W of the system. Setting $v = 0$ in (2.3), we obtain

$$y(k) = \sum_{l=0}^{N-1} W(k, l) u(l), \tag{2.6}$$

with

$$W(k, l) = C G(k, l) B. \tag{2.7}$$

2.2. Stationarity. In contrast with the causal case, where time-invariant state-space models have a time-invariant impulse response, the weighting pattern $W(k, l)$ given by (2.7) is not, in general, a function of the difference $k - l$. TPBVDSs that have this property are called *stationary*.

THEOREM 2.1 [15]. *The* TPBVDS (1.3)-(1.5) *is stationary if and only if*

$$O_s[V_i, E] R_s = O_s[V_i, A] R_s = 0, \tag{2.8a}$$

$$O_s[V_f, E] R_s = O_s[V_f, A] R_s = 0, \tag{2.8b}$$

where $[X, Y]$ *denotes the commutator product of* X *and* Y

$$[X, Y] = XY - YX \tag{2.9}$$

and

$$R_s = [E^{n-1} B A E^{n-2} B \cdots A^{n-1} B], \tag{2.10a}$$

$$O_s^T = [(E^{n-1})^T C^T (A E^{n-2})^T C^T \cdots (A^{n-1})^T C^T]. \tag{2.10b}$$

The matrices R_s and O_s in (2.10) are the *strong reachability* and *strong observability* matrices of the TPBVDS. If they have full rank, the triplets (E, A, B) and (C, E, A) are said, respectively, to be strongly reachable, and strongly observable (see [14]-[15] for a detailed study of the properties of strong and weak reachability and observability). The stationarity conditions (2.8a) and (2.8b) state that V_i and V_f must commute with E and A, except for parts that are either in the left nullspace of R_s or the right nullspace of O_s. Consequently, if R_s and O_s have full rank, i.e., if the TPBVDS is strongly reachable and strongly observable, V_i and V_f must commute with E and A.

It is easily verified that the weighting pattern of a stationary TPBVDS defined over $[0, N]$ is given by

$$W(k)=\begin{cases} CV_iA^{k-1}E^{N-k}B, & 1\leqq k\leqq N, \\ -CV_fE^{-k}A^{N+k-1}B, & 1-N\leqq k\leqq 0. \end{cases} \tag{2.11}$$

2.3. Minimality. Since our goal is to realize shift-invariant acausal weighting patterns with stationary TPBVDSs, we need to be able to determine whether or not a system in this class is minimal. This issue was examined in [15] and [18], leading to the following definition and characterization of minimality.

DEFINITION 2.1. A TPBVDS is *minimal* if its state x has the lowest dimension among all TPBVDSs having the same weighting pattern.

THEOREM 2.2. *The stationary* TPBVDS (1.3)-(1.5) *is minimal if and only if*

(2.12a) $[V_iR_s \quad V_fR_s]$ *has full row rank,*

(2.12b) $\begin{bmatrix} O_sV_i \\ O_sV_f \end{bmatrix}$ *has full column rank,*

(2.12c) $\ker(O_s)\subset \operatorname{im}(R_s)$.

It was also shown in Corollary 5.1 of [15] that Theorem 2.2 implies the following corollary.

COROLLARY. *Let $(C_j, V_i^j, V_f^j, E_j, A_j, N)$ with $j=1, 2$ be two minimal and stationary realizations of the same weighting pattern, where $\{E_j, A_j\}$, $j=1, 2$ are in standard form for the same α and β. Then, there exists an invertible matrix T such that*

$$B_2=TB_1, \qquad C_2=C_1T^{-1}, \tag{2.13a}$$

$$O_s^1(V_i^1-T^{-1}V_i^2T)R_s^1=O_s^1(V_f^1-T^{-1}V_f^2T)R_s^1=0, \tag{2.13b}$$

and

$$(A_1-T^{-1}A_2T)R_s^1=(E_1-T^{-1}E_2T)R_s^1=0, \tag{2.13c}$$

$$O_s^1(A_1-T^{-1}A_2T)=O_s^1(E_1-T^{-1}E_2T)=0, \tag{2.13d}$$

where R_s^1 and O_s^1 are the strong reachability and observability matrices for system 1.

2.4. Extendibility. The concept of extendibility was introduced in [15] for stationary TPBVDSs. It was later extended to nonstationary TPBVDSs in [18]. In this paper, we shall consider only the stationary case.

DEFINITION 2.2. The stationary TPBVDS (1.3)-(1.5) is *extendible* (or input-output extendible) if given any interval $[K, L]$ containing $[0, N]$, there exists a stationary TPBVDS over this larger interval with the same dynamics as in (1.3), but with new boundary matrices $V_i(K, L)$ and $V_f(K, L)$ such that the weighting pattern $W_N(k)$ of the original system is the restriction of the weighting pattern $W_{L-K}(k)$ of the new extended system, i.e.,

$$W_N(k)=W_{K-L}(k) \quad \text{for } 1-N\leqq k\leqq N. \tag{2.14}$$

Our characterization of the property of extendibility for stationary TPBVDSs relies on the notion of Drazin inverse of a matrix [19, p. 8].

DEFINITION 2.3. Let F be an arbitrary square matrix, and let T be an invertible real transformation such that

$$F=T\begin{bmatrix} M & 0 \\ 0 & N \end{bmatrix}T^{-1}, \tag{2.15a}$$

where M is invertible and N is nilpotent. For example, the real Jordan form of F has the above structure. The *Drazin inverse* of F is defined as

$$F^D = T\begin{bmatrix} M^{-1} & 0 \\ 0 & 0 \end{bmatrix} T^{-1}. \tag{2.15b}$$

It can be shown that the Drazin inverse is unique and possesses the following properties:

(i) F^D can be expressed as a polynomial of F, so that it commutes with F. Thus, if a subspace is F-invariant, it is also F^D-invariant.

(ii) When F is invertible, $F^D = F^{-1}$.

(iii) If μ is the degree of nilpotency of N, i.e., if $N^{\mu-1} \neq 0$ and $N^\mu = 0$, then for $k \geqq \mu$

$$F^{k+1}F^D = F^k. \tag{2.16}$$

(iv) Let G be any matrix. Then, the condition

$$\ker(F^\mu) \subset \ker(G) \tag{2.17a}$$

is equivalent to

$$GFF^D = G. \tag{2.17b}$$

(v) If $\{E, A\}$ is a regular pencil in standard form, we have [18, pp. 33-34]

$$EE^D + AA^D - EE^DAA^D = I. \tag{2.18}$$

The extendibility property can then be characterized as follows.

THEOREM 2.3 [15]. *A stationary* TPBVDS *is extendible if and only if*

$$O_s(V_i - V_iE^DE)R_s = 0, \tag{2.19a}$$

$$O_s(V_f - V_fA^DA)R_s = 0. \tag{2.19b}$$

From conditions (2.19), by using the E-, A-, E^D-, and A^D-invariance of im (R_s) [15] and the generalized Cayley-Hamilton theorem, it is easy to check that for an extendible stationary TPBVDS, the weighting pattern (2.11) can be rewritten as

$$W(k) = \begin{cases} CV_iE^NE^D(AE^D)^{k-1}B, & 1 \leqq k \leqq N, \\ -CV_fA^NA^D(EA^D)^{-k}B, & 1-N \leqq k \leqq 0. \end{cases} \tag{2.20}$$

Given an extendible stationary TPBVDS over $[0, N]$ with weighting pattern $W_N(k)$, it is of interest to ask whether it is possible to extend this TPBVDS in a consistent way over intervals of increasing lengths, i.e., so that this progressive extension process gives rise to a unique extended weighting pattern $W(k)$ defined for all k. A procedure to achieve this objective is given by Theorem 2.4.

THEOREM 2.4. *An extendible stationary* TPBVDS *admits extendible extensions over any interval. Furthermore, the weighting pattern of these extendible extensions is unique.*

Proof. Given an extendible stationary TPBVDS $(C, V_i, V_f, E, A, B, N)$, consider the TPBVDS $(C, \tilde{V}_i, \tilde{V}_f, E, A, B, M)$ defined over an interval of length $M > N$, with

$$\tilde{V}_i = V_iE^N(E^D)^M, \qquad \tilde{V}_f = V_fA^N(A^D)^M. \tag{2.21}$$

It is easy to check that this new TPBVDS is in normalized form, and by using the E-, A-, E^D-, and A^D-invariance of im (R_s), that it is stationary and extendible. According to (2.20), its weighting pattern can be expressed as

$$\tilde{W}_M(k) = \begin{cases} C\tilde{V}_iE^ME^D(AE^D)^{k-1}B, & 1 \leqq k \leqq M, \\ -C\tilde{V}_fA^MA^D(EA^D)^{-k}B, & 1-M \leqq k \leqq 0. \end{cases} \tag{2.22}$$

Substituting (2.21) inside (2.22), and noting from $N > n$ that we have $E^{M+N}(E^D)^M = E^N$ and $A^{M+N}(A^D)^M = A^N$, we find

$$\tilde{W}_M(k) = \begin{cases} CV_iE^NE^D(AE^D)^{k-1}B, & 1 \leqq k \leqq M, \\ -CV_fA^NA^D(EA^D)^{-k}B, & 1-M \leqq k \leqq 0. \end{cases} \tag{2.23}$$

This implies

$$\tilde{W}_M(k) = W_N(k) \quad \text{for } 1-N \leqq k \leqq N, \tag{2.24}$$

so that the TPBVDS $(C, \tilde{V}_i, \tilde{V}_f, E, A, B, M)$ specified by (2.21) is an extension of $(C, V_i, V_f, E, A, B, N)$.

To prove the uniqueness of the extended weighting pattern $\tilde{W}_M(k)$, observe that if $\tilde{W}_M(k)$ is the weighting pattern of an arbitrary extendible extension of TPBVDS (1.3)-(1.5) to an interval of length $M > N$, it can be expressed as (2.22), and satisfies (2.24), so that it is uniquely specified on $[1-N, N]$. Since $N > n$, by applying the standard Cayley-Hamilton theorem to matrices AE^D and EA^D in (2.22), we see that $\tilde{W}_M(k)$ is also uniquely specified on $[N+1, M]$ and $[1-M, -N]$. □

Thus, we can associate to an extendible system a sequence of extendible systems over progressively larger intervals, and with consistent weighting patterns. In this way, we can construct an infinite weighting pattern, called the *extended weighting pattern* of the system, which is such that the weighting pattern of the system and of all its extensions are restrictions of this extended weighting pattern.

From (2.23), the extended weighting pattern of an extendible stationary TPBVDS (1.3)-(1.5) is given by

$$W(k) = \begin{cases} C(V_iE^N)E^D(AE^D)^{k-1}B, & k > 0, \\ -C[I-(V_iE^N)]A^D(EA^D)^{-k}B, & k \leqq 0, \end{cases} \tag{2.25}$$

where we have taken into account the normalization (2.2).

3. Internal description of a weighting pattern. The matrix V_iE^N specifies entirely the effect of the boundary conditions on the extended weighting pattern $W(k)$ given by (2.25). This motivates the introduction of the following concept.

DEFINITION 3.1. Let $(C, V_i, V_f, E, A, B, N)$ be a stationary and extendible TPBVDS. Then P is a *decomposition matrix* of this system if

$$O_sPR_s = O_s(E^NV_i)R_s. \tag{3.1}$$

The motivation for calling P a decomposition matrix is that the extended weighting pattern (2.25) can be expressed as

$$W(k) = \begin{cases} CPE^D(AE^D)^{k-1}B, & k > 0, \\ -C(I-P)A^D(EA^D)^{-k}B, & k \leqq 0. \end{cases} \tag{3.2}$$

Thus, if the identity matrix is decomposed into P and $I-P$, the matrices P and $I-P$ appear as parameters of the causal and anticausal parts of $W(k)$. Also, by using (2.8), (2.19), (3.1), and the fact that im (R_s) and ker (O_s) are E- and A-invariant, it is easy to check that a decomposition matrix P satisfies

$$O_s(PA-AP)R_s = O_s(PE-EP)R_s = 0, \tag{3.3a}$$

$$O_s(P-PEE^D)R_s = 0, \tag{3.3b}$$

$$O_s[(I-P)-(I-P)AA^D]R_s = 0. \tag{3.3c}$$

As is clear from Definition 3.1, one particular choice of decomposition matrix is

$$P = V_iE^N. \tag{3.4}$$

This choice is not unique in general. If P is a decomposition matrix, so is $P+Q$, where Q is any matrix such that O_sQR_s equals zero.

The expression (3.2) for the extended weighting pattern $W(k)$ motivates the introduction of the following concept.

DEFINITION 3.2. A five-tuple (C, P, E, A, B) is said to be an *internal description* of the acausal weighting pattern $W(k)$ if it satisfies (3.2) and (3.3), and if $\{E, A\}$ is in standard form. Furthermore, (C, P, E, A, B) is *minimal* if it has the smallest dimension among all internal descriptions of $W(k)$.

Given an acausal weighting pattern $W(k)$, a possible procedure for constructing a minimal, extendible, stationary TPBVDS $(C, V_i, V_f, E, A, B, N)$ that admits $W(k)$ as extended weighting pattern consists therefore in dividing the realization problem into two steps. First, find a minimal internal description (C, P, E, A, B) of $W(k)$. Next, given a finite interval $[0, N]$, find some appropriate boundary matrices V_i and V_f such that the corresponding TPBVDS is extendible and stationary, and such that P is a decomposition matrix associated to these matrices. The following result guarantees the validity of this two-step realization approach.

THEOREM 3.1. *Consider a weighting pattern $W(k)$ with internal description (C, P, E, A, B). Then, for any interval length N, there exists matrices V_i and V_f such that the* TPBVDS $(C, V_i, V_f, E, A, B, N)$ *is normalized, extendible, stationary, and has $W(k)$ as its extended weighting pattern. P is a decomposition matrix of the* TPBVDS $(C, V_i, V_f, E, A, B, N)$. *Furthermore, this* TPBVDS *is minimal if and only if the internal description (C, P, E, A, B) of $W(k)$ is minimal.*

Proof. Let

$$V_i = P(E^D)^N + \sigma X(\sigma E^N + A^N)^{-1}, \tag{3.5a}$$

$$V_f = (I-P)(A^D)^N + X(\sigma E^N + A^N)^{-1}, \tag{3.5b}$$

where

$$X = I - PEE^D - (I-P)AA^D = (I-P)EE^D + PAA^D - EE^DAA^D, \tag{3.6}$$

and where σ is any scalar such that $\sigma E^N + A^N$ is invertible. The second equality in (3.6) is a consequence of identity (2.18). Relations (3.5)-(3.6) specify a TPBVDS $(C, V_i, V_f, E, A, B, N)$. By direct calculation, it is easy to check that V_i and V_f satisfy the normalization (2.2), and that the stationarity and extendibility conditions (2.8) and (2.19) for $(C, V_i, V_f, E, A, B, N)$ are implied, respectively, by the relations (3.3a) and (3.3b)-(3.3c) for (C, P, E, A, B). Noting that

$$O_sXR_s = 0, \tag{3.7}$$

we can also verify that the extended weighting pattern (2.28) associated to $(C, V_i, V_f, E, A, B, N)$ is equal to the weighting pattern $W(k)$ given by (3.2). Finally, taking (3.7) and (3.3b) into account, the matrix V_i given by (3.5a) satisfies

$$O_sV_iE^NR_s = O_sPE^DER_s = O_sPR_s, \tag{3.8}$$

so that P is a decomposition matrix of $(C, V_i, V_f, E, A, B, N)$.

If the internal description (C, P, E, A, B) is not minimal, there exists an internal description (C', P', E', A', B') of smaller dimension, and the above construction yields an extendible stationary TPBVDS realizing $W(k)$, of smaller dimension than $(C, V_i, V_f, E, A, B, N)$, thus showing that this last TPBVDS is not minimal. Conversely, if the TPBVDS $(C, V_i, V_f, E, A, B, N)$ given by (3.5)-(3.6) is not minimal, we can find a lower-dimensional stationary TPBVDS $(C', V_i', V_f', E', A', B', N)$ that is a minimal realization of $W(k)$ over $[0, N]$. According to Corollary 5.2 of [15], this TPBVDS must be extendible and has $W(k)$ for extended weighting pattern. Then $P' = V_i' E'^N$ is a decomposition matrix for this lower-dimensional realization, thus showing that the internal description (C, P, E, A, B) is not minimal. □

Given an internal description (C, P, E, A, B) of the weighting pattern $W(k)$, the following result shows that it is possible to characterize the minimality of this internal description directly, without invoking minimality conditions for an associated TPBVDS.

THEOREM 3.2. *The internal description* (C, P, E, A, B) *of* $W(k)$ *is minimal if and only if*

(3.9a) $R_w = [R_s \quad PR_s]$ *has full row rank,*

(3.9b) $O_w = \begin{bmatrix} O_s \\ O_s P \end{bmatrix}$ *has full column rank,*

(3.9c) $\ker(O_s) \subset \operatorname{im}(R_s)$.

Proof. According to Theorem 3.1, we can associate to (C, P, E, A, B) an extendible stationary TPBVDS $(C, V_i, V_f, E, A, B, N)$, which is minimal if and only if (C, P, E, A, B) is minimal. This TPBVDS is minimal if and only if conditions (2.12) are satisfied. Thus, we need only to show that conditions (2.12) and (3.9) are equivalent. Suppose that conditions (2.12) are satisfied, but (3.9a) is not. Then, there exists $v \neq 0$ such that

$$v^T R_s = 0, \tag{3.10a}$$

$$v^T P R_s = 0. \tag{3.10b}$$

But from (3.10a) and (2.12c) we can conclude that v^T must belong to the row space of O_s. From (3.10b), we find

$$v^T P R_s = v^T V_i E^N R_s = 0. \tag{3.11}$$

Combining (3.10a) and (3.11) yields

$$v^T V_f A^N R_s = 0. \tag{3.12}$$

Since the system is extendible, we have

$$v^T V_i E E^D R_s = v^T V_i R_s, \tag{3.13a}$$

$$v^T V_f A A^D R_s = v^T V_f R_s. \tag{3.13b}$$

But, since R_s is E- and A-invariant, the range spaces of $E^N R_s$ and $A^N R_s$ coincide, respectively, with the ranges of $EE^D R_s$ and $AA^D R_s$. Combining (3.11)-(3.13), we obtain

$$v^T V_i R_s = v^T V_f R_s = 0, \tag{3.14}$$

which contradicts the assumption that (2.12a) is satisfied. Thus, (3.9a) is implied by (2.12). A similar argument can be used to show that (3.9b) is implied by (2.12).

To prove the converse, assume that (3.9) is satisfied and (2.12a) is not. Then, there exists $v \neq 0$ such that

$$v^T V_i R_s = v^T V_f R_s = 0, \tag{3.15}$$

which because of the E- and A-invariance of R_s implies

$$v^T V_i E^N R_s = v^T V_f A^N R_s = 0. \tag{3.16}$$

This in turn implies

$$v^T R_s = 0, \tag{3.17}$$

so that according to (3.9c) v^T belongs to the row space of O_s. Thus,

$$v^T V_i E^N R_s = v^T P R_s = 0. \tag{3.18}$$

But (3.17) and (3.18) contradict (3.9a). Consequently, (2.12a) is implied by (3.9). A similar argument shows that (3.9) implies (2.12b), thus proving the theorem. □

In the following, by analogy with the weak reachability and observability matrices that were introduced in [14] and [15], to characterize the concepts of weak reachability and observability for a TPBVDS $(C, V_i, V_f, E, A, B, N)$, the matrices R_w and O_w appearing in (3.9a) and (3.9b) will be called the *weak reachability* and *weak observability* matrices of the internal description (C, P, E, A, B). As will be shown below, these two matrices play a key role in the construction of a minimal internal description of the weighting pattern $W(k)$.

Theorem 3.2 implies that two minimal internal descriptions of a weighting pattern $W(k)$ can be related as follows.

COROLLARY. *Consider two minimal internal descriptions $(C_j, P_j, E_j, A_j, B_j)$, with $j=1, 2$, of the same weighting pattern $W(k)$, which are in standard form for the same α and β. Then, there exists an invertible matrix T such that relations* (2.13a), (2.13c)-(2.13d), *and*

$$O_s^1(P_1 - T^{-1}P_2T)R_s^1 = 0 \tag{3.19}$$

are satisfied.

Proof. According to Theorem 3.1, we can construct two minimal TPBVDSs $(C, V_i^j, V_f^j, E, A, B, N)$ associated to the two given internal descriptions of $W(k)$. Then, there exists an invertible matrix T such that relations (2.13) are satisfied. Consequently, the strong reachability and observability matrices of systems 1 and 2 are related through

$$R_s^2 = TR_s^1, \qquad O_s^2 = O_s^1 T^{-1}. \tag{3.20}$$

From (2.13b) and (3.19), we can deduce that

$$O_s^1 V_i^1 R_s^1 = O_s^2 V_i^2 R_s^2, \tag{3.21}$$

which implies

$$O_s^1 V_i^1 E_1^N R_s^1 = O_s^2 V_i^2 E_2^N R_s^2, \tag{3.22}$$

or equivalently,

$$O_s^1 P_1 R_s^1 = O_s^2 P_2 R_s^2, \tag{3.23}$$

which proves (3.19). □

4. Realizability conditions and separable realization. In § 3, we have reduced the minimal TPBVDS realization problem to the following problem. Given an infinite weighting pattern $W(k)$, find a minimal internal description (C, P, E, A, B) of $W(k)$.

4.1. Realizability conditions. As a first step, we characterize the weighting patterns that admit a finite-dimensional internal description.

THEOREM 4.1. *A sequence $W(k)$ admits a finite-dimensional internal description if and only if there exists scalars a_i, $1 \leqq i \leqq n_f$ and b_i, $1 \leqq i \leqq n_b$ such that*

$$W(n_f + l) = \sum_{i=1}^{n_f} a_i W(n_f + l - i) \quad \textit{for all } l > 0, \tag{4.1a}$$

$$W(-n_b + l) = \sum_{i=1}^{n_b} b_i W(-n_b + l + i) \quad \textit{for all } l \leqq 0. \tag{4.1b}$$

Proof. Necessity is shown by applying the standard Cayley–Hamilton theorem to matrices AE^D and EA^D in (3.2). To prove sufficiency, consider the decomposition

$$W(k)=W_f(k)+W_b(k), \tag{4.2a}$$

$$W_f(k)=W(k)1(k-1), \qquad W_b(k)=W(k)1(-k), \tag{4.2b}$$

of $W(k)$ into its causal and anticausal parts, where $1(k)$ denotes the unit step function, i.e.,

$$1(k)=\begin{cases}1 & \text{for } k\geqq 0,\\ 0 & \text{for } k<0.\end{cases}$$

Conditions (4.1a) and (4.1b) imply that $W_f(k)$ and $W_b(k)$ can be realized by finite-dimensional causal and anticausal systems, respectively. Let (C_f, A_f, B_f) and (C_b, A_b, B_b) be such realizations, so that

$$W_f(k)=C_fA_f^{k-1}B_f \quad \text{for } k>0, \tag{4.3a}$$

$$W_b(k)=C_bA_b^{-k}B_b \quad \text{for } k\leqq 0. \tag{4.3b}$$

Then, it is clear that

$$C=[C_f - C_b], \qquad P=\begin{bmatrix} I & 0\\ 0 & 0\end{bmatrix}, \tag{4.4a}$$

$$E=\begin{bmatrix} I & 0\\ 0 & A_b\end{bmatrix}, \quad A=\begin{bmatrix} A_f & 0\\ 0 & I\end{bmatrix}, \quad B=\begin{bmatrix} B_f\\ B_b\end{bmatrix} \tag{4.4b}$$

is an internal description of $W(k)$, so that the theorem is proved. □

4.2. Separable realization. Let us continue to analyze the realization obtained by decomposing the weighting pattern $W(k)$ into causal and anticausal components, and then constructing minimal realizations for each of these components separately. Given a finite interval $[0, N]$, the internal description (4.4) yields the following extendible stationary TPBVDS realization of $W(k)$:

$$\begin{bmatrix} I & 0\\ 0 & A_b\end{bmatrix}x(k+1)=\begin{bmatrix} A_f & 0\\ 0 & I\end{bmatrix}x(k)+\begin{bmatrix} B_f\\ B_b\end{bmatrix}u(k), \tag{4.5a}$$

$$\begin{bmatrix} I & 0\\ 0 & 0\end{bmatrix}x(0)+\begin{bmatrix} 0 & 0\\ 0 & I\end{bmatrix}x(N)=\begin{bmatrix} v_f\\ v_b\end{bmatrix}, \tag{4.5b}$$

$$y(k)=[C_f - C_b]x(k). \tag{4.5c}$$

An interesting feature of this realization is that it consists of two decoupled subsystems, which propagate, respectively, in the forward and backward directions. An extendible stationary TPBVDS with this structure is called *separable.* Also, observe that in (4.5b) the boundary matrices satisfy $V_i=P$ and $V_f=I-P$, regardless of the interval length. Thus, in the separable case, there is no real distinction between internal descriptions and minimal TPBVDS realizations.

Unfortunately, the separable realization (4.5) is not always minimal, as can be seen from the following example, which is an adaptation of an example presented in [20] for boundary-value systems with standard nondescriptor dynamics.

Example 4.1. Consider the weighting pattern

$$W(k)=\begin{cases}1, & k\geqq 1,\\ \frac{1}{2}, & k\leqq 0.\end{cases} \tag{4.6}$$

Its separable realization takes the form

$$x_f(k+1)=x_f(k)+u(k), \qquad x_f(0)=v_f, \tag{4.7a}$$

$$x_b(k)=x_b(k+1)+\tfrac{1}{2}u(k), \qquad x_b(N)=v_b, \tag{4.7b}$$

$$y(k)=y_f(k)+y_b(k). \tag{4.7c}$$

However, this realization is not minimal, since W admits also the following one-dimensional realization:

$$x(k+1)=x(k)+u(k), \tag{4.8a}$$

$$y(k)=\tfrac{1}{2}x(k), \tag{4.8b}$$

$$2x(0)-x(N)=v. \tag{4.8c}$$

In this case, the reason we can realize both the causal and anticausal parts of W with a single one-dimensional system is that they have both the same mode, $\sigma=1$. □

In general, when the causal and anticausal parts of W share a common mode, the realization approach consisting in constructing separately minimal realizations of the causal and anticausal parts of W does not yield a minimal realization.

4.3. Summable weighting patterns. Nevertheless, the separable realization (4.5) turns out to be minimal for the class of weighting patterns $W(k)$ that are *summable*, i.e., such that

$$\sum_{k=-\infty}^{\infty} \|W(k)\| < \infty, \tag{4.9}$$

where $\|\cdot\|$ denotes an arbitrary matrix norm. This class of weighting patterns is important, since it corresponds to BIBO stable systems.

THEOREM 4.2. *When the extended weighting pattern $W(k)$ is summable, the separable* TPBVDS *realization* (4.4)-(4.5) *is strongly reachable and observable, and is therefore minimal.*

Proof. Since $W(k)$ is summable, its causal and anticausal parts $W_f(k)$ and $W_b(k)$ are also summable. This implies that the matrices A_f and A_b appearing in the minimal realizations of W_f and W_b are stable, i.e., their eigenvalues are located inside the unit circle. Consider now the matrix

$$[sE-tA\,|\,B]=\begin{bmatrix} sI-tA_f & 0 & B_f \\ 0 & sA_b-tI & B_b \end{bmatrix}. \tag{4.10}$$

It is shown in Theorem 4.1 of [14] that if this matrix has full rank for $(s,t)\neq(0,0)$, the system is strongly reachable. But since A_f and A_b are stable, the eigenmodes of $sI-tA_f$ and sA_b-tI are such that $s/t<1$ and $s/t>1$, respectively, so that these two matrix pencils do not have any common eigenmode. Furthermore, since the state-space realizations (C_f, A_f, B_f) and (C_b, A_b, B_b) are minimal, the submatrices

$$[sI-tA_f\,|\,B_f] \quad \text{and} \quad [sA_b-tI\,|\,B_b]$$

have full rank. This implies that the matrix (4.10) has full rank, so that TPBVDS (4.5) is strongly reachable. By a similar argument, we can show that

$$\begin{bmatrix} sE-tA \\ C \end{bmatrix}=\begin{bmatrix} sI-tA_f & 0 \\ 0 & sA_b-tI \\ C_f & C_b \end{bmatrix} \tag{4.11}$$

has full rank and that the TPBVDS (4.5) is strongly observable. According to Theorem 3.2, the TPBVDS (4.5) is therefore minimal. □

In the remainder of the paper, we will focus our attention on the general case where $W(k)$ is not summable. In this case, minimal realizations are usually not separable. To obtain a minimal realization, two approaches are possible. One method consists in starting from a nonminimal TPBVDS realization, say the separable realization (4.5), and then using the procedure described in [15] for removing the components of this TPBVDS that are not weakly reachable, not weakly observable, or simultaneously not strongly reachable and observable. An alternative realization approach, that we shall follow here, relies on the introduction of a new transform, the (s, t)-transform, and on formulating the realization problem as a state-space representation problem in the (s, t) domain.

5. The (s, t)-transform and state-space representation of rational matrices. One difficulty associated with the use of the z-transform for analyzing discrete-time descriptor systems is that since the dynamics of such systems are singular, infinite frequencies cannot be handled in the same way as other frequencies [21]. This motivates the introduction of the transform

$$H(s,t)=\sum_{k=-\infty}^{\infty} H(k)t^{k-1}/s^{k}. \tag{5.1}$$

It can be expressed in terms of the standard z-transform $H(z)$ as

$$H(s,t)=H(s/t)/t. \tag{5.2}$$

Relation (5.2) shows that the z-transform can be obtained from the (s, t)-transform simply by replacing (s, t) by $(z, 1)$, and conversely, the (s, t)-transform is obtained from the z-transform by replacing z with s/t, and dividing the result by t. From (5.2), we see also that when $H(s, t)$ exists, it has a particular type of homogeneity and is strictly proper in (s, t) in the sense that

$$\lim_{c\to\infty} H(cs,ct)=\lim_{c\to\infty} H(s,t)/c=0. \tag{5.3}$$

Note, however, that it is not necessarily strictly proper in s and t separately, so that the corresponding z-transform may not be proper.

In the following, we shall restrict our attention to the case when $H(z)$ and $H(s, t)$ are *rational*. Then, from (5.2), we see that the numerator and denominator polynomials of all entries of $H(s, t)$ are *homogeneous*, i.e., each such polynomial has the form

$$p(s,t)=\sum_{i=0}^{d} p_i s^{d-i}t^{i},$$

where d is the degree of p. Furthermore, from (5.3), we see also that the relative degree in s and t of all entries of $H(s, t)$, i.e., the difference between the denominator and numerator degrees is exactly one. Thus, the transformation (5.2) has the effect of transforming rational matrices $H(z)$, proper or not, into strictly proper homogeneous rational matrices in the two variables s and t with relative degree one. The analysis of this paper will focus exclusively on this specific class of rational matrices. Note that the idea of studying the structure at infinity of rational matrices in z through the introduction of a homogenizing transform is not totally new. It has been considered, for example, in [22, pp. 158-162, 182-187] and [23].

5.1. Formulation of the realization problem. In the causal case, the z-transform plays an important role in the solution of the minimal realization problem. Specifically, given a causal weighting pattern $W(k)$, the minimal realization problem is equivalent to finding matrices (C, A, B) of minimal dimension such that the z-transform $W(z)$ admits the *state-space representation*

$$W(z) = C(zI - A)^{-1}B. \tag{5.4}$$

For the case of acausal weighting patterns, the situation is more complex. If (C, P, E, A, B) is an internal description of the weighting pattern $W(k)$, and if $W_f(k)$ and $W_b(k)$ are the causal and anticausal parts of $W(k)$, the (s, t)-transforms of $W_f(k)$ and $W_b(k)$ can be expressed as

$$W_f(s,t) = \sum_{k=1}^{\infty} CPE^D(AE^D)^{k-1}Bt^{k-1}/s^k = CPE^D(sI - tAE^D)^{-1}B, \tag{5.5a}$$

$$\begin{aligned} W_b(s,t) &= \sum_{k=-\infty}^{0} -C(I-P)A^D(EA^D)^kBt^{k-1}/s^k \\ &= C(I-P)A^D(sEA^D - tI)^{-1}B. \end{aligned} \tag{5.5b}$$

Then, we use the matrix identities [19, p. 80]

$$(sE - tA)^{-1} = E^D(sI - tAE^D)^{-1} - A^D(I - EE^D)\sum_{k=0}^{\mu_E - 1}(sEA^D)^k/t^{k+1}, \tag{5.6a}$$

$$(sE - tA)^{-1} = A^D(sEA^D - tI)^{-1} + E^D(I - AA^D)\sum_{k=0}^{\mu_A - 1}(tAE^D)^k/s^{k+1}, \tag{5.6b}$$

where μ_E and μ_A denote the indices of the nilpotent parts of E and A, respectively. Taking into account the properties (3.3a)-(3.3c) of the decomposition matrix P, we obtain

$$W_f(s,t) = CP(sE - tA)^{-1}B = C(sE - tA)^{-1}PB, \tag{5.7a}$$

$$W_b(s,t) = C(I-P)(sE - tA)^{-1}B = C(sE - tA)^{-1}(I-P)B. \tag{5.7b}$$

Note that $W_f(s, t)$ and $W_b(s, t)$ do not have, in general, the same regions of convergence. However, by analytic continuation, it is possible to extend their domains of definition to the whole plane while using the same notation. This yields the three representations:

$$W(s,t) = W_f(s,t) + W_b(s,t) = C(sE - tA)^{-1}B, \tag{5.8}$$

$$H_r(s,t) = [\,W_f(s,t)\; W_b(s,t)] = C(sE - tA)^{-1}[PB\;(I-P)B], \tag{5.9}$$

$$H_o(s,t) = \begin{bmatrix} W_f(s,t) \\ W_b(s,t) \end{bmatrix} = \begin{bmatrix} CP \\ C(I-P) \end{bmatrix}(sE - tA)^{-1}B. \tag{5.10}$$

Since the specification of an acausal weighting pattern $W(k)$ is equivalent to the specification of $W_f(s, t)$ and $W_b(s, t)$, we see from (5.8)-(5.10) that the construction of an internal description (C, P, E, A, B) of $W(k)$ can be expressed as a state-space realization problem for rational matrices in s and t. However, in contrast to the causal case, the need to specify P and to achieve minimality implies that we must, in general, obtain state-space representations for the *three* rational matrices $W(s, t)$, $H_r(s, t)$, and $H_b(s, t)$, instead of a single rational matrix for causal systems. Furthermore, since we are considering acausal systems, the computation of any of these state-space representations requires an extension of known state-space realization techniques. We consider this problem first in the next section.

5.2. State-space representations of homogeneous rational matrices in s and t. The above discussion motivates the following *minimal state-space representation problem.* Given an homogeneous rational matrix function $H(s, t)$ of relative degree one, find matrices (K, D, F, G) of lowest possible dimension such that

$$H(s, t) = K(sD - tF)^{-1}G. \tag{5.11}$$

This problem is the counterpart of the minimal state-space representation problem for a strictly proper rational matrix $H(z)$, where we seek to find matrices (K, F, G) of smallest size such that

$$H(z) = K(zI - F)^{-1}G. \tag{5.12}$$

The difference between (5.11) and (5.12) is that, as was noted earlier, the one-dimensional rational transfer function $H(z) = H(z, 1)$ associated to (5.11) is not necessarily proper, so that the representation (5.12) is not applicable to this case.

An important feature of the minimal representation (5.12) is that it is unique up to a similarity transform. For the minimal representation (5.11), even if we impose the additional requirement that $\{D, F\}$ should be in standard form, i.e., that there exists α and β such that

$$\alpha D + \beta F = I, \tag{5.13}$$

the matrices (K, D, F, G) are not unique. To ensure uniqueness, α and β must be chosen a priori. In the causal case, i.e., when $H(z)$ is strictly proper, this was done implicitly in (5.12) by forcing D to be equal to I, which corresponds to selecting $\alpha = 1$ and $\beta = 0$. For the more general case that we consider here, any pair (α, β) is acceptable as long as

$$H(\alpha, -\beta) < \infty. \tag{5.14}$$

This last condition can be viewed as an extension of the condition $H(\infty) < \infty$ for proper transfer functions.

THEOREM 5.1. *A matrix function $H(s, t)$ admits a state-space representation* (5.11) *if and only if it is rational, homogeneous in s and t, and with relative degree one. Under these conditions, if (α, β) is a pair of scalars such that $H(\alpha, -\beta)$ exists, $H(s, t)$ admits a unique minimal representation, up to a similarity transform, satisfying* (5.11) *and* (5.13). *The dimension r of this minimal realization, i.e., the size of D and F, is given by*

$$r = d(H(\alpha z, 1 - \beta z)), \tag{5.15}$$

where $d(\cdot)$ denotes the usual McMillan degree, and where $H(\alpha z, 1 - \beta z)$ is a strictly proper rational matrix in z.

Proof. If $H(s, t)$ admits a representation of the form (5.11), it is clear that it must be rational, homogeneous in s and t, and of relative degree one. To prove sufficiency, we need to construct such a representation. Let α and β be such that $H(\alpha, -\beta)$ exists. Then, consider the rational matrix $H(\alpha z, 1 - \beta z)$. This matrix is strictly proper in z because

$$\lim_{z \to \infty} H(\alpha z, 1 - \beta z) = \lim_{z \to \infty} H(\alpha, -\beta)/z = 0. \tag{5.16}$$

It can therefore be realized as

$$H(\alpha z, 1 - \beta z) = K(zI - F)^{-1}G. \tag{5.17}$$

Now, assume that $\alpha \neq 0$ (otherwise, reverse the roles of D and F), and let

$$w = \frac{\alpha}{\alpha t + \beta s}, \qquad z = \frac{s}{\alpha t + \beta s}. \tag{5.18}$$

In this case

$$s = \frac{\alpha z}{w}, \qquad t = \frac{1 - \beta z}{w}, \tag{5.19}$$

which implies that

$$H(s, t) = wH(\alpha z, 1 - \beta z) = wK(zI - F)^{-1}G = K(sD - tF)^{-1}G, \tag{5.20}$$

with

$$D = \frac{I - \beta F}{\alpha}. \tag{5.21}$$

Since there is a one-to-one correspondence between the representation (5.17) of $H(\alpha z, 1 - \beta z)$ and the representation (5.20) of $H(s, t)$ with D given by (5.21), the dimension and uniqueness properties of these two representations are the same. This implies that minimal state-space representations of $H(s, t)$ satisfying (5.20) and (5.21) are related by a similarity transform, and have a dimension r equal to the McMillan degree of $H(\alpha z, 1 - \beta z)$. □

COROLLARY. *The state-space representation* (5.11), (5.13) *is minimal if and only if* (D, F, G) *is strongly reachable and* (K, D, F) *is strongly observable. Furthermore, the dimension of a minimal state-space representation is equal to the rank of the Hankel matrix* O_sR_s, *where* O_s *and* R_s *are the strong observability and reachability matrices associated, respectively, to* (K, D, F) *and* (D, F, G).

Proof. It can be assumed without loss of generality that $\alpha \neq 0$ in (5.13). Then, the representation (5.11), (5.13) of $H(s, t)$ is minimal if and only if the representation (5.17) of $H(\alpha z, 1 - \beta z)$ is minimal, or equivalently if and only if (K, F) is observable and (F, G) is reachable, where observability and reachability are defined here in the sense of causal systems. Since $\alpha \neq 0$, this is equivalent to requiring that (K, D, F) and (D, F, G) are strongly observable, and strongly reachable, respectively (see [14, Thm. 4.1]).

It was also shown in Theorem 5.1 that the dimension r of a minimal state-space representation is equal to the McMillan degree of $H(\alpha z, 1 - \beta z)$. But according to the realization theory of causal systems, this McMillan degree is equal to the rank of the Hankel matrix

$$\bar{H} = \overline{OR}, \tag{5.22}$$

where $\bar{O}$ and $\bar{R}$ are the observability and reachability matrices associated to the pairs (K, F) and (F, G), respectively. But with $\alpha \neq 0$, the nullspace of $\bar{O}$ coincides with that of O_s, and the range of $\bar{R}$ with that of R_s. This implies that the rank of $\bar{H}$ is equal to that of O_sR_s, thus proving the corollary. □

One relatively unsatisfactory aspect of Theorem 5.1 is that the dimension r of a minimal state-space representation of $H(s, t)$ is characterized in terms of the McMillan degree of the one-dimensional rational matrix $H(\alpha z, 1 - \beta z)$, and not directly in terms of $H(s, t)$. It turns out that it is possible to characterize r directly from $H(s, t)$ by extending the concept of McMillan degree as follows.

DEFINITION 5.1. Given a homogeneous and strictly proper rational matrix $H(s, t)$ in s and t, the *McMillan degree* of $H(s, t)$ is defined as the degree of the least common multiple of the denominators of all minors of $H(s, t)$.

Then, we have Theorem 5.2.

THEOREM 5.2. *If $H(s, t)$ is realizable, i.e., if it is homogeneous of relative degree one, the dimension of a minimal state-space representation of $H(s, t)$ is equal to its McMillan degree.*

Proof. Consider the minimal representation

$$H(s, t) = K(sD - tF)^{-1}G. \tag{5.23}$$

Without loss of generality, it can be assumed that the pencil $sD - tF$ is in Weierstrass canonical form (see [24, p. 28]), so that

$$K = [K_1 \quad K_2], \quad D = \begin{bmatrix} D_1 & 0 \\ 0 & N \end{bmatrix}, \quad F = \begin{bmatrix} F_1 & 0 \\ 0 & F_2 \end{bmatrix}, \quad G = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}, \tag{5.24}$$

where N is nilpotent and D_1 and F_2 are invertible. The rational matrix

$$H_1(s, t) = K_1(sD_1 - tF_1)^{-1}G_1 \tag{5.25}$$

can then be expressed as

$$H_1(s, t) = K_1(zD_1 - F_1)^{-1}G_1/t = \bar{H}_1(z)/t, \tag{5.26}$$

where $z = s/t$. Since there is a one-to-one correspondence between $H_1(s, t)$ and $\bar{H}_1(z)$, the dimensions of minimal representations of these two rational matrices must be equal. But, $\bar{H}_1(z)$ is a strictly proper rational matrix in z, so that the dimension of its minimal representation is equal to its McMillan degree, i.e., to the degree of the least common multiple $a_1(z)$ of the denominators of all minors of $\bar{H}_1(z)$. Also, since D_1 is invertible, t is not a factor of the denominator of any of the entries, and thus of any of the minors of $H_1(t, s)$. Let $p_1(s, t)$ denote the least common multiple of the denominators of the minors of $H_1(s, t)$. Since t is not a factor of $p_1(s, t)$, the degree of $p_1(s, t)$ is just the degree in z of $p_1(z, 1) = a_1(z)$. This shows that the degree of $p_1(s, t)$ equals the McMillan degree of $\bar{H}_1(z)$, which is in turn equal to the dimension of D_1 and F_1.

For the second block of the representation (5.24), we proceed similarly. Let

$$H_2(s, t) = K_2(sN - tF_2)^{-1}G_2, \tag{5.27}$$

and denote by $p_2(s, t)$ the least common multiple of the denominators of the minors of $H_2(s, t)$. Since F_2 is invertible, s is not a factor of $p_2(s, t)$. This implies that the degree of $p_2(s, t)$ is just the degree in t of $p_2(1, t)$, which, by analogy with the previous case, is just the dimension of N and F_2. Also, since N is nilpotent and (N, F_2) is in standard form, we have

$$p_2(s, t) = \det(sN - tF_2) = at^{n_2}, \tag{5.28}$$

where a is a constant, and n_2 the dimension of N and F_2.

Noting that

$$H(s, t) = H_1(s, t) + H_2(s, t) \tag{5.29}$$

and the fact that $p_1(s, t)$ and $p_2(s, t)$ have no common factors, we can easily deduce that the least common multiple $p(s, t)$ of the denominators of the minors of H satisfies

$$p(s, t) = p_1(s, t)p_2(s, t). \tag{5.30}$$

The degree of $p(s, t)$ is therefore equal to the sum of the dimensions of the blocks of (5.24), which is the dimension of D and F. □

Example 5.1. Consider the sequence

$$H(k)=\begin{cases}-1, & k=0,\\ 1, & k=1,\\ 0 & \text{otherwise.}\end{cases} \tag{5.31}$$

Its (s, t)- and z-transforms are, respectively,

$$H(s, t)=\frac{1}{s}-\frac{1}{t}, \tag{5.32a}$$

$$H(z)=-1+\frac{1}{z}. \tag{5.32b}$$

Already we can see the advantage of using the (s, t)-transform: $H(s, t)$ has one mode at zero and one at infinity, where $H(z)$ has only a pole at $z=0$.

From Theorem 5.2, we see that the dimension of a minimal representation, simply select $\alpha=\beta=1$, and perform the realization

$$H(z, 1-z)=\frac{1}{z}-\frac{1}{1-z}=[1 \quad 1]\left(zI-\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}\right)^{-1}\begin{bmatrix}1\\ 1\end{bmatrix}, \tag{5.33}$$

which implies that

$$K=[1 \quad], \quad D=\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}, \quad F=\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}, \quad G=\begin{bmatrix}1\\ 1\end{bmatrix}. \tag{5.34}$$

6. Minimal realization.

In § 5, it was shown that the specification of an internal description (C, P, E, A, B) of a weighting pattern $W(k)$ yields the three state-space representations (5.8)-(5.10) for the rational matrices $W(s, t)$, $H_r(s, t)$, and $H_o(s, t)$. This suggests that the construction of a minimal internal description of $W(k)$ can be formulted as a state-space representation problem in the (s, t)-domain. It turns out that the link existing between minimal state-space representations of rational matrices and minimal internal descriptions is less direct than for causal systems, since an internal description (C, P, E, A, B) can be minimal, even though *none* of the state-space representations (5.8)-(5.10) is minimal.

6.1. Dimension of a minimal realization.

THEOREM 6.1. *The dimension n of a minimal internal description of $W(k)$ is given by*

$$n=\omega+\rho-\tau, \tag{6.1}$$

where if $d(\cdot)$ denotes the generalized McMillan degree introduced in Definition 5.1,

$$\omega=d(H_r(s, t)), \quad \rho=d(H_o(s, t)), \quad \tau=d(W(s, t)). \tag{6.2}$$

Proof. Let (C, P, E, A, B) be a minimal internal description of $W(k)$. Then, $W(s, t)$, $H_r(s, t)$, and $H_o(s, t)$ admit state-space representations of the form (5.8)-(5.10), and from the corollary of Theorem 5.1, ω, ρ, and τ are the ranks of the Hankel matrices O_sR_w, O_wR_s, and O_sR_s, respectively. But, according to the minimality conditions (3.9a)-(3.9b), R_w and O_w have full rank, which implies that ω and ρ are the ranks of the strong observability and reachability matrices O_s and R_s, respectively. From condition (3.9c), we can also deduce that the rank of O_sR_s equals the rank of O_s plus that of R_s minus n, so that

$$\tau=\rho+\omega-n, \tag{6.3}$$

which implies (6.1). □

Example 6.1. Consider the weighting pattern

$$W(k)=\begin{cases} a^k & k\geqq 1, \\ ba^k, & k<1, \end{cases} \tag{6.4}$$

where a and b are scalar parameters with $a<1$. Using Theorem 4.1, it is straightforward to check that $W(k)$ is realizable. From Theorem 6.1, we find that the dimension of a minimal internal description of $W(k)$ is given by

$$\begin{aligned} n&=d\Bigg(\Bigg[\frac{a}{s-at}\ \frac{ab}{s-at}\Bigg)\Bigg]+d\left(\begin{bmatrix}\dfrac{a}{s-at}\\ \dfrac{-ab}{s-at}\end{bmatrix}\right)-d\left(\frac{(1-b)a}{s-at}\right)\\ &=\begin{cases}1+1-1=1 & \text{for } b\neq 1,\\ 1+1-0=2 & \text{for } b=1.\end{cases}\end{aligned} \tag{6.5}$$

When $b\neq 1$, a minimal internal description of $W(k)$ is

$$C=\frac{a}{1-b},\quad P=\frac{1}{1-b},\quad E=1,\quad A=a,\quad B=1. \tag{6.6}$$

The causal and anticausal parts $W_f(s, t)$ and $W_b(s, t)$ of W have the same pole, namely $s/t=a$, which explains why they can be realized with a single eigenmode. The resulting TPBVDS realization is strongly reachable, strongly observable, and nonseparable.

When $b=1$, a minimal internal description of $W(k)$ is

$$C=[a\quad a],\quad P=\begin{bmatrix}1&0\\0&0\end{bmatrix},\quad E=I,\quad A=aI,\quad B=\begin{bmatrix}1\\1\end{bmatrix}. \tag{6.7}$$

This separable realization is not strongly reachable, and is not strongly observable. Note that in the realization (6.6), the system matrices tend to ∞ as $b\to 1$. Thus, $b=1$ can be viewed as a singularity in the sense that the dimension of a minimal internal description of W is two only when b is exactly equal to one. □

6.2. Minimal realization procedure. One interesting aspect of Theorem 6.1 is that as an intermediate step in the evaluation of the dimension n of a minimal internal description of $W(k)$, we obtain ω and ρ, which are, respectively, the ranks of the strong observability and reachability matrices of a minimal internal description. This observation leads to the following procedure for constructing a minimal internal description of W.

Step 1. Construct the minimal state-space representations

$$H_r(s, t)=[W_f(s, t)\,W_b(s, t)]=\bar{C}(s\bar{E}-t\bar{A})^{-1}[\bar{B}_f\quad \bar{B}_b], \tag{6.8}$$

$$H_o(s, t)=\begin{bmatrix}W_f(s, t)\\ W_b(s, t)\end{bmatrix}=\begin{bmatrix}\tilde{C}_f\\ \tilde{C}_b\end{bmatrix}(s\tilde{E}-t\tilde{A})^{-1}\tilde{B}, \tag{6.9}$$

where if α and β are such that $W_f(\alpha, -\beta)$ and $W_b(\alpha, -\beta)$ are defined, the pairs $\{\bar{E}, \bar{A}\}$ and $\{\tilde{E}, \tilde{A}\}$ satisfy the normalization condition (2.1) for the same α and β. Since the representations (6.8) and (6.9) are both minimal, the sizes of the matrices $\{\bar{E}, \bar{A}\}$ and $\{\tilde{E}, \tilde{A}\}$ are equal, respectively, to ω and ρ.

Step 2. Let

$$\bar{B} = \bar{B}_f + \bar{B}_b, \qquad \tilde{C} = \tilde{C}_f + \tilde{C}_b. \tag{6.10}$$

From (6.8)-(6.9), we find

$$\begin{aligned} W(s, t) &= W_f(s, t) + W_b(s, t) \\ &= \bar{C}(s\bar{E} - t\bar{A})^{-1}\bar{B} = \tilde{C}(s\tilde{E} - t\tilde{A})^{-1}\tilde{B}, \end{aligned} \tag{6.11}$$

so that $(\bar{C}, \bar{E}, \bar{A}, \bar{B})$ and $(\tilde{C}, \tilde{E}, \tilde{A}, \tilde{B})$ are two state-space representations, in general nonminimal, of $W(s, t)$. The minimality of representations (6.8) and (6.9) implies that $(\bar{C}, \bar{E}, \bar{A}, \bar{B})$ and $(\tilde{C}, \tilde{E}, \tilde{A}, \tilde{B})$ are, respectively, strongly observable and strongly reachable. By decomposing these two representations into strongly reachable/unreachable, and strongly observable/unobservable components, respectively, we obtain

$$\bar{C} = [\bar{C}_1 \quad \bar{C}_2], \quad \bar{E} = \begin{bmatrix} \bar{E}_1 & \bar{E}_2 \\ 0 & \bar{E}_4 \end{bmatrix}, \quad \bar{A} = \begin{bmatrix} \bar{A}_1 & \bar{A}_2 \\ 0 & \bar{A}_4 \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} \bar{B}_1 \\ 0 \end{bmatrix}, \tag{6.12}$$

and

$$\tilde{C} = [0 \quad \tilde{C}_2], \quad \tilde{E} = \begin{bmatrix} \tilde{E}_1 & \tilde{E}_2 \\ 0 & \tilde{E}_4 \end{bmatrix}, \quad \tilde{A} = \begin{bmatrix} \tilde{A}_1 & \tilde{A}_2 \\ 0 & \tilde{A}_4 \end{bmatrix}, \quad \tilde{B} = \begin{bmatrix} \tilde{B}_1 \\ \tilde{B}_2 \end{bmatrix}. \tag{6.13}$$

In the following, it will be assumed that the representations (6.8) and (6.9) are in the coordinate systems corresponding to (6.12) and (6.13), respectively.

Step 3. From (6.11), we find that

$$W(s, t) = \bar{C}_1(s\bar{E}_1 - t\bar{A}_1)^{-1}\bar{B}_1 = \tilde{C}_2(s\tilde{E}_4 - t\tilde{A}_4)^{-1}\tilde{B}_2, \tag{6.14}$$

where the representations $(\bar{C}_1, \bar{E}_1, \bar{A}_1, \bar{B}_1)$ and $(\tilde{C}_2, \tilde{E}_4, \tilde{A}_4, \tilde{B}_2)$ are both strongly reachable and observable. This implies that they must be related by a similarity transformation, i.e., there exists a matrix T such that

$$\bar{C}_1 = \tilde{C}_2 T^{-1}, \quad \bar{E}_1 = T\tilde{E}_4 T^{-1}, \quad \bar{A}_1 = T\tilde{A}_4 T^{-1}, \quad \bar{B}_1 = T\tilde{B}_2. \tag{6.15}$$

The matrix T is given by

$$T = \bar{M}_s \tilde{M}_s^T (\tilde{M}_s \tilde{M}_s^T)^{-1}, \tag{6.16}$$

where $\bar{M}_s$ and $\tilde{M}_s$ denote, respectively, the strong reachability matrices of $(\bar{E}_1, \bar{A}_1, \bar{B}_1)$ and $(\tilde{E}_4, \tilde{A}_4, \tilde{B}_2)$. Furthermore, since the representations (6.14) are minimal, the matrices $\bar{E}_1$, $\bar{A}_1$, $\tilde{E}_4$ and $\tilde{A}_4$ have dimension τ, where τ is given by (6.2), and consequently, the blocks $\{\bar{E}_4, \bar{A}_4\}$, and $\{\tilde{E}_1, \tilde{A}_1\}$ in the decompositions (6.12) and (6.13) have respective dimensions $\omega - \tau$ and $\rho - \tau$.

Step 4. The matrices C, E, A, and B of a minimal internal description are now selected as

$$E = \begin{bmatrix} \tilde{E}_1 & \tilde{E}_2 T^{-1} & * \\ 0 & \bar{E}_1 & \bar{E}_2 \\ 0 & 0 & \bar{E}_4 \end{bmatrix}, \qquad A = \begin{bmatrix} \tilde{A}_1 & \tilde{A}_2 T^{-1} & * \\ 0 & \bar{A}_1 & \bar{A}_2 \\ 0 & 0 & \bar{A}_4 \end{bmatrix}, \tag{6.17a}$$

$$C = [0 \quad \bar{C}_1 \quad \bar{C}_2], \qquad B = \begin{bmatrix} \tilde{B}_1 \\ \bar{B}_1 \\ 0 \end{bmatrix}, \tag{6.17b}$$

where * indicates an arbitrary block entry. The role of the similarity transformation T is to guarantee that the component which is common to state-space representations (6.8) and (6.9) is expressed in the same coordinate system. Note that (5.17) corresponds to a four part Kalman decomposition of (C, E, A, B) into strongly reachable/unreachable and observable/unobservable parts, where according to (3.9c), there is no strongly unreachable and unobservable component, since the internal description that we are constructing must be minimal.

By using this last observation, we can immediately conclude from (6.11) that

$$W(s,t)=C(sE-tA)^{-1}B. \tag{6.18}$$

If we denote

$$B_f=\begin{bmatrix} * \\ \bar{B}_f \end{bmatrix}, \qquad C_f=[\tilde{C}_f \quad *]\begin{bmatrix} I & 0 & 0 \\ 0 & T^{-1} & 0 \\ 0 & 0 & I \end{bmatrix}, \tag{6.19}$$

from (6.8) and (6.9), it is also easy to check that

$$W_f(s,t)=C(sE-tA)^{-1}B_f=C_f(sE-tA)^{-1}B. \tag{6.20}$$

Expanding $W_f(s,t)$ in power series of $s-\alpha$ and $t+\beta$ in the vicinity of $(s,t)=(\alpha,-\beta)$, noting that $\alpha E+\beta A=I$, and matching the coefficients of $(s-\alpha)^i(t+\beta)^j$ for all i, j in (6.20) yields

$$O_sR_s^f=O_s^fR_s, \tag{6.21}$$

where R_s^f and O_s^f denote the strong reachability and observability matrices associated respectively to (E, A, B_f) and (C_f, E, A).

Step 5. The matrix P is then obtained by solving the equation

$$O_sPR_s=O_sR_s^f. \tag{6.22}$$

The existence of a solution is guaranteed by identity (6.21), which shows that the row and column spaces of the matrix on the right side of (6.22) are spanned by O_s and R_s, respectively. The solution of (6.22) is generally not unique, since we can add to any solution P a matrix Q such that $O_sQR_s=0$, i.e., a matrix of the form

$$Q=\begin{bmatrix} * & * & * \\ 0 & 0 & * \\ 0 & 0 & * \end{bmatrix}. \tag{6.23}$$

We must now prove that the matrices (C, P, E, A, B) given by (6.17) and (6.22) specify an internal description of $W(k)$. This requires showing that the state-space representation identities (5.9)-(5.10) are satisfied, as well as properties (3.3). The relation

$$O_sPR_s=O_sR_s^f=O_s^fR_s, \tag{6.24}$$

implies

$$CPR_s=CR_s^f, \qquad O_sPB=O_s^fB, \tag{6.25}$$

so that from the Cayley-Hamilton theorem and (6.20), we have

$$CP(sE-tA)^{-1}B=C(sE-tA)^{-1}PB=W_f(s,t). \tag{6.26}$$

When combined with (6.18), this yields the representations (5.9)-(5.10). To prove relations (3.3a), we use (6.24) and the fact that the reachability matrices R_s and R^f_s, and observability matrices O_s and O^f_s are constructed from the same matrices E and A. Then, from the Cayley-Hamilton theorem, there exist matrices K_E and K_A which satisfy

$$ER_s = R_s K_E, \qquad AR_s = R_s K_A, \tag{6.27a}$$

$$ER^f_s = R^f_s K_E, \qquad AR_s = R^f_s K_A, \tag{6.27b}$$

i.e., the same matrices K_E and K_A can be used to characterize the E- and A-invariance of the range spaces of both R_s and R^f_s. Similarly, the E- and A-invariance of the nullspaces of O_s and O^f_s can be characterized by a single pair of matrices. Taking this feature into account in (6.24), it can be checked easily that the constraints (3.3a) are satisfied. To prove relations (3.3b) and (3.3c), we use identities (5.6a)-(5.6b). Substituting (5.6a) inside (6.26), and noting that the weighting pattern $W_f(k)$ is causal, we find

$$CPA^D(I-EE^D)(EA^D)^k B = 0 \tag{6.28}$$

for $0 \leqq k \leqq \mu_E - 1$. Expressing the pencil $\{E, A\}$ in Weierstrass canonical form, it is then easy to check that (6.28) is equivalent to (3.3b). Similarly, to derive (3.3c), we substitute (5.6b) inside the state-space representation

$$W_b(s,t) = C(I-P)(sE - tA)^{-1}B \tag{6.29}$$

and use the fact that $W_b(k)$ is an anticausal weighting pattern. Thus, (C, P, E, A, B) is an internal description of $W(k)$. Since its dimension n obeys (6.1), it is *minimal.*

Example 6.2. Let

$$W(k) = \begin{cases} 0, & k = 1, \\ -1, & k \neq 1. \end{cases} \tag{6.30}$$

Then

$$W_f(s,t) = \frac{-t}{s(s-t)}, \qquad W_b(s,t) = \frac{1}{s-t}, \tag{6.31}$$

and according to Theorem 6.1, the dimension of a minimal internal description of $W(k)$ is

$$n = 2 + 2 - 1 = 3. \tag{6.32}$$

Since $\omega = \rho = 2$, we can also conclude that the minimal internal description is neither strongly reachable nor strongly observable. To obtain a minimal description, the first step is to construct the minimal state-space representations

$$[W_f \quad W_b] = \begin{bmatrix} \dfrac{-t}{s(s-t)} & \dfrac{1}{s-t} \end{bmatrix} = [1 \quad 1]\left(sI - t\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}\right)^{-1} \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}, \tag{6.33a}$$

$$\begin{bmatrix} W_f \\ W_b \end{bmatrix} = \begin{bmatrix} \dfrac{-t}{s(s-t)} \\ \dfrac{1}{s-t} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & 0 \end{bmatrix}\left(sI - t\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}\right)^{-1} \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \tag{6.33b}$$

which satisfy the normalization condition (2.1) with $\alpha = 1$, $\beta = 0$. This yields

$$\bar{B} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \tilde{C} = [1 \quad 0]. \tag{6.34}$$

In this case, we can select $T=1$, and

$$C=[0 \quad 1 \quad 1], \quad E=I, \quad A=\begin{bmatrix} 1 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad B=\begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \tag{6.35}$$

where * denotes an arbitrary entry. Finally, by solving (6.22) we find

$$P=\begin{bmatrix} * & * & * \\ 0 & 1 & * \\ 1 & 0 & * \end{bmatrix}. \tag{6.36}$$

The above realization procedure can be simplified significantly if the minimal internal description is either strongly observable or strongly reachable, i.e., if the integers ω and ρ in (6.2) satisfy either $\omega = n$ or $\rho = n$.

Strongly observable case $(\omega = n)$. In this case, only the state-space representation (6.8) is needed, and we can select $(C, E, A, B) = (\bar{C}, \bar{E}, \bar{A}, \bar{B})$. Also, since O_s has full rank, (6.24) for P reduces to

$$PR_s = R_s^f. \tag{6.37}$$

Strongly reachable case $(\rho = n)$. Then, only the representation (6.9) is needed, and we can select $(C, E, A, B) = (\tilde{C}, \tilde{E}, \tilde{A}, \tilde{B})$. Furthermore, (6.24) for P becomes

$$O_s P = O_s^f. \tag{6.38}$$

The previous realization procedure, or its simplification for the strongly observable and reachable cases, is of interest only when it yields a minimal internal description which is not separable, since in the separable case, the realization of § 4 is minimal. the following result provides a test for determining whether a weighting pattern admits a separable minimal description.

THEOREM 6.2. *$W(k)$ has a separable minimal realization if and only if the minimal dimension n given by* (6.1) *satisfies*

$$n = d(W_f(s, t)) + d(W_b(s, t)). \tag{6.39}$$

Proof. If we construct two minimal realizations of W_f and W_b, and combine them to realize $W(k)$ as shown in (4.4)-(4.5), we obtain a description of dimension $d(W_f(s, t)) + d(W_b(s, t))$. This description will therefore be minimal if and only if (6.39) is satisfied, where n is given by (6.1). □

7. Conclusions. In this paper, the minimal TPBVDS realization problem for acausal shift-invariant weighting patterns has been examined. By restricting our attention to extendible stationary TPBVDSs, it was shown that the minimal TPBVDS realization problem is equivalent to the problem of finding a minimal internal description for the weighting pattern $W(k)$ of interest. Introducing the (s, t) transform and characterizing minimal state-space representations of homogeneous rational matrices in (s, t), a frequency-domain approach was developed for finding the dimension of a minimal internal description, and for constructing such a description.

Since the assumption that the weighting pattern $W(k)$ is shift-invariant is restrictive, particularly for acausal systems, it would be of interest to extend the above theory to the nonstationary case. Also, we have limited our attention here to deterministic systems. Since there exists a complete and elegant stochastic realization theory for causal systems [25]-[27], it is natural to ask whether a similar theory can be developed for acausal stochastic systems. For the Gaussian case, some preliminary stochastic realization results have been presented in [28] for boundary value systems with standard nondescriptor dynamics, and in [18] for TPBVDSs.

REFERENCES

[1] B. L. Ho and R. E. Kalman, *Effective construction of linear state-variable models from input-output functions*, in Proc. Third Allerton Conference, 1966, pp. 449-459.

[2] D. C. Youla, *The synthesis of linear dynamical systems from prescribed weighting patterns*, SIAM J. Appl. Math. 14 (1966), pp. 527-549.

[3] L. M. Silverman, *Representation and realization of time-variable linear systems*, Ph.D. dissertation, Department of Electrical Engineering, Columbia University, New York, 1966.

[4] R. E. Kalman, P. L. Falb, and M. A. Arbib, *Topics in Mathematical System Theory*, McGraw-Hill, New York, 1969.

[5] F. L. Lewis, *Descriptor systems: Decomposition into forwards and backwards subsystems*, IEEE Trans. Automat. Control, 29 (1984), pp. 167-170.

[6] D. G. Luenberger, *Dynamic systems in descriptor form*, IEEE Trans. Automat. Control, 22 (1977), pp. 312-321.

[7] ———, *Time-invariant descriptor systems*, Automatica, 14 (1978), pp. 473-480.

[8] ———, *Boundary recursion for descriptor variable systems*, IEEE Trans. Automat. Control, 34 (1989), pp. 287-292.

[9] A. J. Krener, *Boundary value linear systems*, Astérisque, 75-76 (1980), pp. 149-165.

[10] ———, *Acausal realization theory, Part* 1: *linear deterministic systems*, SIAM J. Control Optim., 25 (1987), pp. 499-525.

[11] I. Gohberg and M. A. Kaashoek, *On minimality and stable minimality of time-varying linear systems with well-posed boundary conditions*, Internat. J. Control, 43 (1986), pp. 1401-1411.

[12] I. Gohberg, M. A. Kaashoek, and L. Lerer, *Minimality and irreducibility of time-invariant linear boundary-value systems*, Internat. J. Control, 44 (1986), pp. 363-379.

[13] I. Gohberg and M. A. Kaashoek, *Similarity and reduction for time-varying linear systems with well-posed boundary conditions*, SIAM J. Control Optim., 24 (1986), pp. 961-978.

[14] R. Nikoukhah, A. S. Willsky, and B. C. Levy, *Boundary-value descriptor systems: well-posedness, reachability and observability*, Internat. J. Control, 46 (1987), pp. 1715-1737.

[15] ———, *Reachability, observability and minimality for shift-invariant two-point boundary-value descriptor systems*, Circuits Systems Signal Process., 8 (1989), pp. 313-340.

[16] R. Nikoukhah, B. C. Levy, and A. S. Willsky, *Stability, stochastic stationarity and generalized Lyapunov equations for two-point boundary-value descriptor systems*, IEEE Trans. Automat. Control, 34 (1989), pp. 1141-1152.

[17] R. Nikoukhah, *System theory for two-point boundary-value descriptor systems*, M.S. thesis, Department of Electrical Engineering and Computer Science, and Report LIDS-TH-1559, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1986.

[18] ———, *A deterministic and stochastic theory for two-point boundary-value descriptor systems*, Ph.D. thesis, Department of Electrical Engineering and Computer Science, and Report LIDS-TH-1820, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, 1988.

[19] S. L. Campbell, *Singular Systems of Differential Equations*, Research Notes in Mathematics, No. 40, Pitman, San Francisco, CA, 1980.

[20] I. Gohberg and M. A. Kaashoek, *Various minimalities for systems with boundary conditions and integral operators*, in Modelling, Identification and Robust Control, C. I. Byrnes and A. Lindquist, eds., North-Holland, Amsterdam, the Netherlands, 1986, pp. 181-196.

[21] G. Verghese and T. Kailath, *Rational matrix structure*, IEEE Trans. Automat. Control, 26 (1981), pp. 434-439.

[22] B. C. Levy, 2-D *polynomial and rational matrices, and their applications for the modeling of* 2-D *dynamical systems*, Report M735-11, Information Systems Laboratory, Stanford University, Stanford, CA, 1981.

[23] S. Tan and J. Vandewalle, *Novel theory for polynomial and rational matrices, Parts* 1 *and* 2, Internat. J. Control, 48 (1988), pp. 545-576.

[24] F. R. Gantmacher, *The Theory of Matrices*, Vol. 2, Chelsea, New York, 1960.

[25] H. Akaike, *Markovian representation of stochastic processes by canonical variables*, SIAM J. Control, 13 (1975), pp. 162-173.

[26] A. LINDQUIST AND G. PICCI, *Realization theory for multivariate stationary Gaussian processes*, SIAM J. Control Optim., 23 (1985), pp. 809-857.

[27] G. RUCKEBUSCH, *Théorie géométrique de la représentation Markovienne*, Thèse de doctorat d'état, Universite de Paris VI, Paris, France, 1980.

[28] A. J. KRENER, *Reciprocal processes and the stochastic realization problem for acausal systems*, in Modelling, Identification and Robust Control, C. I. Byrnes and A. Lindquist eds., North-Holland, Amsterdam, the Netherlands, 1986, pp. 197-209.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 445-456, March 1991

011

THE DYNAMIC PROGRAMMING EQUATION FOR THE TIME-OPTIMAL CONTROL PROBLEM IN INFINITE DIMENSIONS*

VIOREL BARBU†

Abstract. The existence and uniqueness of a viscosity solution for the Bellman equation associated with the time-optimal control problem for a semilinear evolution equation in Hilbert space is provided. Applications to time-optimal control problems governed by parabolic equations are given.

Key words. time-optimal control problem, minimal time function, Hamilton-Jacobi equation, viscosity solution

AMS(MOS) subject classifications. 35F20, 35F25, 49B10, 49C05

1. Introduction. In this paper, we study the time-optimal control problem

$$\inf\{T;\ \exists u\in \mathcal{U},\ y'+Ay+Fy=Bu \text{ in } (0;T);\ y(0)=y_0,\ y(T)=0\} \tag{1.1}$$

in a real Hilbert space H, where A is a linear self-adjoint positive-definite operator in H, F is a nonlinear maximal monotone (single-valued) operator, and

$$\mathcal{U}=\{u\in L^2_{\text{loc}}(R^+;\ U);\ u(t)\in K \text{ a.e. } t>0\}. \tag{1.2}$$

Here U is a real Hilbert space and K is a closed convex and bounded subset of U; $0\in K$. B is a linear continuous operator from U to H and B^* is its adjoint.

Throughout in the sequel we shall denote by $|\cdot|$ and $(\cdot,\cdot)$ the norm and the scalar product of H, respectively. The norm of U will be denoted $|\cdot|_U$ and the scalar product $\langle\cdot,\cdot\rangle$. Further assume that the operator $A+F$ is maximal monotone. Then it is well known (see, for instance, [1], [5]) that for every $y_0\in\overline{D(A+F)}$ (the closure of the domain $D(A+F)$ of the operator $A+F$) the Cauchy problem

$$\begin{aligned} &y'+Ay+Fy=Bu \quad \text{in } (0,T),\\ &y(0)=y_0 \end{aligned} \tag{1.3}$$

has a unique weak solution $y=y(t,y_0,u)\in C([0,T];H)$. (We have denoted by $C([0,T];H)$ the space of all H-valued continuous functions on $[0,T]$.)

For every $y_0\in\overline{D(A+F)}$ let us denote by $\phi(y_0)$ the value of problem (1.1), i.e.,

$$\phi(y_0)=\inf\{T;\ \exists u\in\mathcal{U},\ y(T,y_0,u)=0\}. \tag{1.4}$$

The function $\phi:\overline{D(A+F)}\to[0,\infty]$ is called *the minimal time function* associated to problem (1.1) and formally it is a solution to the stationary Hamilton-Jacobi equation (the Bellman equation)

$$h(-B^*D\phi(y))+(Ay+Fy,D\phi(y))=1 \quad \forall y\in D(A+F)\backslash\{0\}, \tag{1.5}$$

$$\phi(0)=0,\quad \phi\geqq 0 \quad \text{in } \overline{D(A+F)}. \tag{1.6}$$

Here $h:H\to R$ is the support function of K, i.e.,

$$h(p)=\sup\{\langle p,u\rangle;\ u\in K\} \quad \forall p\in H. \tag{1.7}$$

We will prove here under appropriate assumptions on the evolution system (1.3) that the minimal time function ϕ is the unique weakly continuous viscosity solution

* Received by the editors September 20, 1989; accepted for publication (in revised form) March 29, 1990.

† Department of Mathematics, University of Iaşi, 6600 Iaşi, Romania.

to (1.5), which satisfies conditions (1.6). This result has been established in the framework of the theory of viscosity solutions in infinite dimensions developed by Crandall and Lions [8]-[11] (see also [12]) and could be viewed as an illustration of this theory on time-optimal control problems. The verification of the general assumptions on distributed control processes governed by semilinear parabolic equations contains much of the substance of this paper. For some related finite-dimensional results we refer to [4]. We might expect to find significant applications of the theory of viscosity solutions to other classes of optimal control problems.

Recently, Tătaru [17] has introduced a general notion of viscosity solution for infinite-dimensional Hamilton–Jacobi equations with nonlinear unbounded terms which cover equations of the form (1.5).

2. Existence and uniqueness of viscosity solutions. We begin by formulating the main assumptions under which (1.5) will be studied.

(i) A is a linear maximal monotone self-adjoint operator in H with compact resolvent. Denote by $D(A)$ the domain of A and by V the space $D(A^{1/2})$ endowed with the norm $\|x\| = (|x|^2 + |A^{1/2}x|^2)^{1/2}$.

(ii) F is a maximal monotone single-valued operator of the form $F = \partial j$ where $j : H \to]-\infty, +\infty]$ is a lower semicontinuous convex function. Moreover, $F0 = 0$, $\overline{D(F)} = H$ and

$$(2.1) \qquad (Ay, Fy) \geqq 0 \quad \forall y \in D(A) \cap D(F) \neq \emptyset.$$

By ∂j we have denoted the subdifferential of function j.

(iii) The operator $(I + A)^{-1}F$ is locally Lipschitz from H to V, i.e., for every $r > 0$ there is $L(r) > 0$ such that

$$(2.2) \qquad \|(I + A)^{-1}(Fy - Fz)\| \leqq L(r)|y - z| \quad \text{for } |y|, |z| \leqq r.$$

(iv) The minimal time function ϕ is finite and weakly continuous on $\overline{D(A+F)} = H$.

It follows by assumptions (2.1) that $A + F$ is maximal monotone in $H \times H$ and $\overline{D(A+F)} = \overline{D(A) \cap D(F)} = H$ (see, e.g., [3, p. 129]). Moreover, $A + F = \partial\tilde{j}$, where $\tilde{j}(y) = \frac{1}{2}|A^{1/2}y|^2 + j(y)$. Recall (see, e.g., [1], [5]) that $-(A+F)$ generates a nonlinear semigroup of contractions $e^{-(A+F)t}$ on $\overline{D(A+F)} = H$.

In particular, assumption (iv) holds if $U = H$, $B = I$ (the unity operator in H) and $K = \{u \in H; |u| \leqq \rho\}$. Indeed in this case we have (see [6])

$$\phi(y_0) - \phi(z_0) \leqq \rho|y_0 - z_0| \quad \forall y_0, z_0 \in \overline{D(A+F)} = H.$$

If $\{y_0^n\}$ is weakly convergent to y_0, then since by assumptions (i) and (ii) the operator $S(t) = e^{-(A+F)t}$ is compact for $t > 0$, we have on a subsequence, again denoted y_0^n, $S(\varepsilon)y_0^n \to S(\varepsilon)y_0$, and therefore

$$\lim_{n\to\infty} \phi(S(\varepsilon)y_0^n) = \phi(S(\varepsilon)y_0).$$

Inasmuch as $\phi(y_0^n) \leqq \phi(S(\varepsilon)y_0^n) + \varepsilon$, for all $\varepsilon > 0$, this yields

$$\limsup_{n\to\infty} \phi(y_0^n) \leqq \varepsilon + \rho|S(\varepsilon)y_0 - y_0| + \phi(y_0).$$

Since ε can be chosen arbitrarily small we may conclude that

$$\limsup_{n\to\infty} \phi(y_0^n) \leqq \phi(y_0).$$

On the other hand, it is readily seen that the function ϕ is weakly lower semicontinuous. Hence $\lim_{n\to\infty} \phi(y_0^n)=\phi(y_0)$ as claimed.

We shall discuss in the next section other situations where these assumptions hold.

Consider the Hamilton-Jacobi equation

$$E(y, \phi, D\phi)+(Ay+Fy, D\phi)=0 \quad \text{in } \Omega \tag{2.3}$$

where Ω is an open subset of H and $E\colon \Omega\times R\times H\to R$ is a continuous function.

DEFINITION 1. A *strong viscosity* solution to (2.3) in Ω is a continuous function $\phi\in C(\Omega)$ provided for all $\psi\in C^1(\Omega)$:

(a) If $\phi-\psi$ attains a local maximum at $y_0\in\Omega$ and $\nabla\psi(y)=\alpha(y)(\nabla\psi_1(y)+\psi_2'(|y|)\operatorname{sgn} y)$ for all $y\in\Omega$, where $\psi_1\in C^1(\Omega)$, $\psi_2\in C^1(R^+)$ and $\alpha\in C(\Omega)$ are such that $A\nabla\psi_1\in C(\Omega)$, $\alpha\geqq 0$ in Ω, $\psi_2'(0)=0$, $\psi_2'(r)>0$ for $r>0$, then

$$E(y_0, \phi(y_0), \nabla\psi(y_0))+\alpha(y_0)(y_0, A\nabla\psi_1(y_0))+\alpha(y_0)((I+A)^{-1}Fy_0, (I+A)\nabla\psi_1(y_0))\leqq 0 \tag{2.4}$$

(b) If $\phi-\psi$ attains a local minimum at $y_0\in\Omega$ and $\nabla\psi(y)=\alpha(y)(\nabla\psi_1(y)-\psi_2'(|y|)\operatorname{sgn} y)$ for all $y\in\Omega$, where ψ_1, ψ_2 and α satisfy the above conditions, then

$$\begin{aligned}E(y_0, \phi(y_0), \nabla\psi(y_0))&+\alpha(y_0)((I+A)^{-1}Fy_0, (I+A)\nabla\psi_1(y_0))\\&+\alpha(y_0)(y_0, A\nabla\psi_1(y_0))\geqq 0.\end{aligned} \tag{2.5}$$

This notion of solution is slightly stronger than that introduced by Crandall and Lions [10], [11] for Hamilton-Jacobi equations in Hilbert spaces. In accordance with this general concept of viscosity solution we say that $\phi\in C(\Omega)$ is a *viscosity solution* to (2.3) provided for all $\psi\in C^1(\Omega)$:

(j) If $\phi-\psi$ attains a local maximum at $y_0\in\Omega$ and $\psi(y)=\psi_1(y)+\psi_2(|y|)$, where $\psi_1\in C^1(\Omega)$ and $\psi_2\in C^1(R^+)$ are such that $A\nabla\psi_1\in C(\Omega)$, $\psi_2(0)=0$, $\psi_2'(r)>0$ for $r>0$, then

$$E(y_0, \phi(y_0), \nabla\psi(y_0))+(y_0, A\nabla\psi_1(y_0))+((I+A)^{-1}Fy_0, (I+A)\nabla\psi_1(y_0))\leqq 0.$$

(jj) If $\phi-\psi$ attains a local minimum at y_0 and $\psi(y)=\psi_1(y)-\psi_2(|y|)$, where ψ_1, ψ_2 are as above, then

$$E(y_0, \phi(y_0), \nabla\psi(y_0))+(y_0, A\nabla\psi_1(y_0))+((I+A)^{-1}Fy_0, (I+A)\nabla\psi_1(y_0))\geqq 0.$$

THEOREM 1. *Let* (i)–(iv) *hold. Then there is a unique weakly continuous strong viscosity solution to* (1.5) *on* $H\backslash\{0\}$ *which satisfies condition* (1.6), *namely, the minimal time function* ϕ.

If F is locally Lipschitz on H, Theorem 1 is implied by the general existence and uniqueness results proved in [11].

Proof. (1) Existence. By the dynamic programming principle we have for all $t>0$

$$\phi(y_0)=\inf\{t+\phi(y(t, y_0, u));\ u\in\mathcal{U}\}. \tag{2.6}$$

Note that for every $y_0\in\overline{D(A+F)}=H, u\in\mathcal{U}$ and $T>0$, $t^{1/2}y(t, y_0, u)\in W^{1,2}([0, T]; H)=\{z\in L^2(0, T; H);\ dz/dt\in L^2(0, T; H)\}$ and

$$\frac{d}{dt}y(t, y_0, u)+Ay(t, y_0, u)+Fy(t, y_0, u)=Bu(t) \quad \text{a.e. } t\in(0, T). \tag{2.7}$$

Moreover, if $u\in W^{1,2}([0, T]; H)$, then $t(dy/dt)\in L^\infty(0, T; H)$ and

$$\frac{d^+}{dt}y(t)+Ay(t)+Fy(t)=Bu(t) \quad \forall t\in(0, T). \tag{2.8}$$

Let $y_0 \in H\backslash\{0\}$ be a local maximum for $\phi - \psi$ where ψ is as in Definition 1. Then, for any $u \in \mathcal{U}$ and $t > 0$,

$$\phi(y_0) - \psi(y_0) \geqq \phi(y(t, y_0, u)) - \psi(y(t, y_0, u))$$

and for $u(t) \equiv u_0 \in K$ it follows by (2.6) and (2.8) that

$$-1 \leqq \lim_{t \to 0} t^{-1} \int_0^t (\nabla\psi(y(s, y_0, u_0)), Bu_0 - Ay(s, y_0, u_0) - Fy(s, y_0, u_0))\, ds$$
$$\leqq (u_0, B^*\nabla\psi(y_0)) - (y_0, A\nabla\psi_1(y_0))\alpha(y_0) - \alpha(y_0)((I+A)^{-1}Fy_0, (I+A)\nabla\psi_1(y_0)),$$

and inasmuch as u_0 is arbitrary in K the conclusion is now

$$h(-B^*\nabla\psi(y_0)) + (y_0, A\nabla\psi_1(y_0))\alpha(y_0) + ((I+A)^{-1}Fy_0, (I+A)\nabla\psi_1(y_0))\alpha(y_0) \leqq 1,$$

as desired.

Assume now that $y_0 \neq 0$ is a local minimum point for $\phi - \psi$ where $\nabla\psi(y) = (\nabla\psi_1(y) - \psi_2'(y)\operatorname{sgn} y)\alpha(y)$. We have

$$\phi(y_0) - \phi(y(t, y_0, u)) \leqq \psi(y_0) - \psi(y(t, y_0, u)) \quad \forall u \in \mathcal{U}, \quad t \in [0, \delta]. \tag{2.9}$$

Let (y^*, u^*) be an optimal pair in problem (1.3), i.e., $y^*(t) = y(t, y_0, u^*)$ and $y^*(\phi(y_0)) = 0$. (The existence of a such a pair is an immediate consequence of the compacity of the level sets $\{y \in H; \frac{1}{2}(Ay, y) + j(y) + |y|^2 \leqq \lambda\}$ and follows by standard arguments.)

Let $y_0^\varepsilon \in D(A) \cap D(F)$ be such that $y_0^\varepsilon \to y_0$ for $\varepsilon \to 0$ and let $y_\varepsilon(t) = y(t, y_0^\varepsilon, u^*)$. Then, we have

$$y_\varepsilon(t) \to y^*(t) \quad \text{uniformly on } [0, T] \text{ in } H$$

and

$$\psi(y_\varepsilon(t)) - \psi(y_0^\varepsilon) = \int_0^t (\nabla\psi(y_\varepsilon(s)), Bu^*(s) - Ay_\varepsilon(s) - Fy_\varepsilon(s))\, ds$$
$$\geqq \int_0^t (\nabla\psi(y_\varepsilon(s)), Bu^*(s))\, ds - \int_0^t (A\nabla\psi_1(y_\varepsilon(s)), y_\varepsilon(s))\, ds$$
$$- \int_0^t ((I+A)\nabla\psi_1(y_\varepsilon(s)), (I+A)^{-1}Fy_\varepsilon(s))\, ds.$$

Letting ε tend to zero, we get

$$\psi(y^*(t)) - \psi(y_0) \geqq \int_0^t (\nabla\psi(y^*(s)), Bu^*(s))\, ds$$
$$- \int_0^t \alpha(y^*(s))(A\nabla\psi_1(y^*(s)), y^*(s))$$
$$- \int_0^t \alpha(y^*(s))((I+A)\nabla\psi_1(y^*(s)), (I+A)^{-1}Fy^*(s))\, ds.$$

Then by (2.6) and (2.7) we infer that

$$1 + \lim_{t \downarrow 0} t^{-1}\left(\int_0^t (B^*\nabla\psi(y^*(s)), u^*(s)\right) ds$$
$$- \int_0^t \alpha(y^*(s))(y^*(s), A\nabla\psi_1(y^*(s)))\, ds$$
$$- \int_0^t \alpha(y^*(s))((I+A)^{-1}Fy^*(s), (I+A)\nabla\psi_1(y^*(s)))\, ds\Big) \leqq 0$$

and therefore

$$\alpha(y_0)(A\nabla\psi_1(y_0), y_0)+\alpha(y_0)((I+A)\nabla\psi_1(y_0), (I+A)^{-1}Fy_0)+h(-B^*\nabla\psi(y_0))\geqq 1$$

as claimed.

(2) Uniqueness. Using a device due to Kružkov [14] (see also [15]), we show first that (1.5) can be reduced via the transformation $\tilde{\phi}=1-e^{-\phi}$ to the equation

$$\tilde{\phi}+h(-B^*D\tilde{\phi})+(Ay+Fy, D\tilde{\phi})=1 \quad \text{in } H\backslash\{0\}. \tag{2.10}$$

Namely, we have the following lemma.

LEMMA 1. *If ϕ is a strong viscosity solution to* (1.5) *in $H\backslash\{0\}$, then the function $\tilde{\phi}=1-e^{-\phi}$ is a viscosity solution to* (2.10).

Proof. Let $\psi\in C^1(H\backslash\{0\})$ be as in condition (j) and let y_0 be a local maximum point for $\tilde{\phi}-\psi$. This yields

$$\phi(y)+\ln(1-e^{\phi(y_0)}(\psi(y)-\psi(y_0)))\leqq\phi(y_0) \quad \forall y\in B(y_0) \tag{2.11}$$

where $B(y_0)=\{y\in H; |y-y_0|\leqq r_0\}$ is a sufficiently small neighborhood of y_0 such that

$$2e^{\phi(y_0)}|\psi(y)-\psi(y_0)|\leqq 1 \quad \forall y\in B(y_0).$$

Consider a function $\tilde{\psi}\in C^1(H\backslash 0)$ such that

$$\tilde{\psi}(y)=\psi(y) \quad \text{for } y\in B(y_0), \qquad |\tilde{\psi}(y)-\psi(y_0)|\leqq\tfrac{2}{3}e^{-\phi(y_0)} \quad \forall y\in H.$$

Denote by χ the function

$$\chi(y)=-\ln(1-e^{\phi(y_0)}(\tilde{\psi}(y)-\tilde{\psi}(y_0)))$$

and since ϕ is a viscosity solution to (1.5) we have by Definition 1

$$e^{\phi(y_0)}h(-B^*\nabla\psi(y_0))+(y_0, A\nabla\psi_1(y_0))\,e^{\phi(y_0)}$$
$$+((I+A)^{-1}Fy_0, (I+A)\nabla\psi_1(y_0))\,e^{\phi(y_0)}\leqq 1$$

i.e.,

$$\tilde{\phi}(y_0)+h(-B^*\nabla\psi(y_0))+((I+A)^{-1}Fy_0, (I+A)\nabla\psi_1(y_0))+(y_0, A\nabla\psi_1(y_0))\leqq 1$$

as claimed.

The case when y_0 is a local minimum point for $\tilde{\phi}-\psi$ and $\psi=\psi_1-\psi_2(|y|)$ can be treated similarly.

According to Lemma 1 to complete the proof of Theorem 1 it suffices to prove that (2.10) has a unique viscosity solution which is weakly continuous and satisfies the conditions

$$\tilde{\phi}(0)=0, \quad 0\leqq\tilde{\phi}\leqq 1 \quad \text{in } H\backslash\{0\}. \tag{2.12}$$

To this end, following a general method to prove uniqueness of the viscosity solution developed in [9] and [10], consider the Hamiltonian

$$H(z, p, q)=h(-B^*p)-h(B^*q)+(Ax+Fx, p)+(Ay+Fy, q), \qquad z=(x, y).$$

Let ϕ_1, ϕ_2 be two viscosity solutions to (2.10) which are weakly continuous on H and satisfy (2.12) and let w be the function

$$w(z)=\phi_1(x)-\phi_2(y), \qquad z=(x, y)\in H\times H.$$

For every $\varepsilon>0$ consider the function $\Phi_\varepsilon: H\times H\to R$ defined by

$$\Phi_\varepsilon(x, y)=\rho(|x|^2+|y|^2)+\varepsilon^{-1}(\varepsilon^{4\omega}+\|x-y\|_*^2)^{1/2\omega} \tag{2.13}$$

where $\rho>0$, $\omega=L^2(1/\sqrt{\rho})+1$, and $\|x\|_*^2=((I+A)^{-1}x, x)$.

Inasmuch as the function $w-\Phi_\varepsilon$ is weakly lower semicontinuous on $H\times H$ there is $z_\varepsilon=(x_\varepsilon, y_\varepsilon)$ such that

$$w(z_\varepsilon)-\Phi_\varepsilon(x_\varepsilon, y_\varepsilon)\geqq w(z)-\Phi_\varepsilon(x, y)\quad \forall(x, y)\in H\times H. \tag{2.14}$$

We have $\Phi_\varepsilon=\rho(|x|^2+|y|^2)+\psi_\varepsilon(x, y)$ where

$$\nabla_x\psi_\varepsilon(x, y)=(\varepsilon\omega)^{-1}(I+A)^{-1}(x-y)(\varepsilon^4+\|x-y\|_*^2)^{1/2\omega-1},$$
$$\nabla_y\psi_\varepsilon(x, y)=-\nabla_x\psi_\varepsilon(x, y).$$

It is readily seen that

$$w+H(z, Dw)=0\quad \text{in } (H\backslash\{0\})\times(H\backslash\{0\})$$

in the viscosity sense. Hence if $x_\varepsilon\neq 0$ and $y_\varepsilon\neq 0$, we have

$$\begin{aligned}w(z_\varepsilon)&+h(-B^*(2\rho x_\varepsilon+\nabla_x\psi_\varepsilon(x_\varepsilon, y_\varepsilon)))-h(B^*(2\rho y_\varepsilon-\nabla_x\psi_\varepsilon(x_\varepsilon, y_\varepsilon)))\\ &+(\varepsilon\omega)^{-1}|x_\varepsilon-y_\varepsilon|^2(\varepsilon^{4\omega}+\|x_\varepsilon-y_\varepsilon\|_*^2)^{1/2\omega-1}\\ &+(\varepsilon\omega)^{-1}((I+A)^{-1}(Fx_\varepsilon-Fy_\varepsilon), x_\varepsilon-y_\varepsilon)(\varepsilon^{4\omega}+\|x_\varepsilon-y_\varepsilon\|_*^2)^{1/2\omega-1}\\ &\leqq(\varepsilon\omega)^{-1}\|x_\varepsilon-y_\varepsilon\|_*^2(\varepsilon^{4\omega}+\|x_\varepsilon-y_\varepsilon\|_*^2)^{1/2\omega-1}.\end{aligned}$$

By assumptions (iii) this yields

$$\begin{aligned}w(z_\varepsilon)-\Phi_\varepsilon(z_\varepsilon)&\leqq C\rho^{1/2}-\varepsilon^{-1}(\varepsilon^{4\omega}+\|x_\varepsilon-y_\varepsilon\|_*^2)^{1/2\omega}\\ &+(\varepsilon\omega)^{-1}\left(L\left(\frac{1}{\sqrt{\rho}}\right)|x_\varepsilon-y_\varepsilon|\|x_\varepsilon-y_\varepsilon\|_*-|x_\varepsilon-y_\varepsilon|^2\right)\\ &\cdot(\varepsilon^{4\omega}+\|x_\varepsilon-y_\varepsilon\|_*^2)^{1/2\omega-1}+(\varepsilon\omega)^{-1}(\varepsilon^{4\omega}+\|x_\varepsilon-y_\varepsilon\|_*^2)^{1/2\omega}\end{aligned} \tag{2.15}$$

because $h(p)-h(q)\leqq C|p-q|$ for all $p, q\in H$.

On the other hand, we see by (2.14) that

$$\rho(|x_\varepsilon|^2+|y_\varepsilon|^2)\leqq\phi_1(x_\varepsilon)\leqq 1. \tag{2.16}$$

Then after some calculation involving (2.15) we get the estimate

$$w(z_\varepsilon)-\Phi_\varepsilon(z_\varepsilon)\leqq C\rho^{1/2}$$

where C is a positive constant independent of ρ and ε. Then by (2.14) we see that

$$\phi_1(x)-\phi_2(x)\leqq C\rho^{1/2}+2\rho|x|^2+\varepsilon\quad \forall x\in H. \tag{2.17}$$

Since ρ and ε can be chosen arbitrarily small, it follows by (2.17) that $\phi_1=\phi_2$ in H.

If $x_\varepsilon=0$, $y_\varepsilon=0$, then again by (2.14) we have

$$\phi_1(x)-\phi_2(x)\leqq 2\rho|x|^2\quad \forall x\in H,$$

and so $\phi_1=\phi_2$. Similarly, if $x_\varepsilon=0$ and $y_\varepsilon\neq 0$ once again using inequality (2.14), we get

$$\rho|y_\varepsilon|^2+\varepsilon^{-1}(\varepsilon^{4\omega}+\|y_\varepsilon\|_*^2)^{1/2\omega}\leqq\varepsilon$$

and therefore $y_\varepsilon=0$.

Let us now assume that $x_\varepsilon\neq 0$ and $y_\varepsilon=0$. Then we have

$$\rho|x_\varepsilon|^2+\varepsilon^{-1}(\varepsilon^{4\omega}+\|x_\varepsilon\|_*^2)^{1/2\omega-1}\leqq\varepsilon+\phi_1(x_\varepsilon)\leqq 1+\varepsilon \tag{2.18}$$

and therefore

$$\|x_\varepsilon\|_*\leqq(\varepsilon(1+\varepsilon))^\omega.$$

Thus if $x_{\varepsilon_n}\neq 0$ and $y_{\varepsilon_n}=0$, for a sequence $\varepsilon_n\to 0$ we have

$$x_{\varepsilon_n}\to 0\quad \text{strongly in } V^* \text{ and weakly in } H$$

and therefore $\phi_1(x_{\varepsilon_n}) \to 0$ (because ϕ_1 is weakly continuous). (We have denoted here by V^* the dual of the space V endowed with the norm $\|\cdot\|_*$.)

Then again by (2.14) we have

$$\phi_1(x) - \phi_2(x) \leqq \phi_1(x_{\varepsilon_n}) + 2\rho|x|^2 + \varepsilon_n \quad \forall x \in H\backslash\{0\}.$$

Hence $\phi_1 = \phi_2$, thereby completing the proof.

3. The time-optimal problem for a semilinear parabolic equation. Consider the distributed control system

$$\begin{aligned} &\frac{\partial y}{\partial t} - \Delta y + \beta(y) = u \quad \text{in } \Omega \times (0, \infty), \\ &y(x, 0) = y_0(x), \quad x \in \Omega, \qquad y = 0 \quad \text{in } \partial\Omega \times (0, \infty) \end{aligned} \tag{3.1}$$

where Ω is a bounded and open subset of R^N with a sufficiently smooth boundary $\partial\Omega$ (for instance, of class C^2), and β is a continuous monotonically increasing function on r such that $\beta(0) = 0$ and

$$|\beta(y) - \beta(z)| \leqq C(1 + |y|^\rho + |z|^\rho)|y - z| \quad \forall y, z \in R \tag{3.2}$$

where $0 \leqq \rho \leqq 2/N$ if $N \geqq 3$, $0 \leqq \rho < 1$ if $N = 2$, and $\rho = 1$ if $N = 1$.

Let $K = \{u \in L^2(\Omega);\ |u(x)| \leqq \alpha \text{ a.e. } x \in \Omega\}$ and let $\phi : L^2(\Omega) \to R^+$ be the minimal time function associated with system (3.1),

$$\phi(y_0) = \inf\{T;\ \exists u \in L^2(0, T; L^2(\Omega)),\ u(t) \in K, \text{a.e. } t \in (0, T);\ y(T, y_0, u) = 0\}. \tag{3.3}$$

Here $y(\cdot, y_0, u) \in C([0, T]; L^2(\Omega))$ is the solution to problem (3.1).

Let us check assumptions (i)-(iv) of § 2 where $H = L^2(\Omega)$, $A = -\Delta$, $D(A) = H_0^1(\Omega) \cap H^2(\Omega)$, $V = H_0^1(\Omega)$, $V^* = H^{-1}(\Omega)$, and $(Fy)(x) = \beta(y(x))$ for almost every $x \in \Omega$, $y \in D(F)$ where

$$D(F) = \{y \in L^2(\Omega);\ \beta(y) \in L^2(\Omega)\}.$$

$H_0^1(\Omega)$, $H^2(\Omega)$, and $H^{-1}(\Omega)$ are Sobolev spaces on Ω.

Since assumptions (i), (ii), are obviously satisfied, we confine ourselves to checking (iii) and (iv).

By the Sobolev imbedding theorem,

$$\|A^{-1}(Fy - Fz)\|_{H_0^1(\Omega)} = \|Fy - Fz\|_{H^{-1}(\Omega)} \leqq C\|\beta(y) - \beta(z)\|_{L^p(\Omega)} \quad \forall y, z \in L^2(\Omega)$$

where $p = 2N/(N+2)$ if $N \geqq 3$, $1 < p \leqq 2$ if $N = 2$, and $p = 1$ if $N = 1$.

Then by (3.2) we see that

$$\|A^{-1}(Fy - Fz)\|_{H_0^1(\Omega)} \leqq C(1 + \|y\|_{L^2(\Omega)}^{2/N} + \|z\|_{L^2(\Omega)}^{2/N})\|y - z\|_{L^2(\Omega)},$$

respectively,

$$\|A^{-1}(Fy - Fz)\|_{H_0^1(\Omega)} \leqq C(1 + \|y\|_{L^2(\Omega)}^{\rho} + \|z\|_{L^2(\Omega)}^{\rho})\|y - z\|_{L^2(\Omega)}$$

for $N = 2$

$$\|A^{-1}(Fy - Fz)\|_{H_0^1(\Omega)} \leqq C(1 + \|y\|_{L^2(\Omega)} + \|z\|_{L^2(\Omega)})\|y - z\|_{L^2(\Omega)}$$

for $N = 1$.

Let us now prove that assumption (iv) holds in the present situation. We shall first prove that any $y_0 \in L^\infty(\Omega)$ can be steered to origin in finite time. To this aim arguing as in [2], consider the feedback system

$$(3.4)\qquad \begin{aligned} &\frac{\partial y}{\partial t} - \Delta y + \beta(y) + \alpha \operatorname{sgn} y = 0 && \text{in } \Omega \times R^+,\\ &y(x,0) = y_0(x) && \text{in } \Omega,\\ &y = 0 && \text{in } \partial\Omega \times R^+ \end{aligned}$$

where $\operatorname{sgn} y = y|y|^{-1}$ if $y \neq 0$ and $\operatorname{sgn} 0 = [-1, 1]$. Recall that problem (3.4) has a unique solution $y \in C([0, T]; L^2(\Omega))$ such that

$$t^{1/2} y \in L^2(0, T; H^2(\Omega) \cap H_0^1(\Omega)), \quad t^{1/2} \frac{\partial y}{\partial t} \in L^2(0, T; L^2(\Omega)) \quad \forall t > 0.$$

Consider the function $z = \|y_0\|_{L^\infty(\Omega)} - \alpha t$ and note that

$$(3.5)\qquad \begin{aligned} &\frac{\partial z}{\partial t} - \Delta z + \alpha \operatorname{sgn} z \ni 0 && \text{in } \Omega \times (0, \alpha^{-1}\|y_0\|_{L^\infty(\Omega)}),\\ &z(x,0) = \|y_0\|_{L^\infty(\Omega)} && \text{in } \Omega,\\ &z \geqq 0 && \text{in } \partial\Omega \times (0, \alpha^{-1}\|y_0\|_{L^\infty(\Omega)}). \end{aligned}$$

Subtracting (3.4) and (3.5) and multiplying by $(y-z)^+$, we get after some calculation that

$$\frac{d}{dt}\|(y-z)^+\|^2_{L^2(\Omega)} \leqq 0 \quad \text{a.e. } t \in (0, \alpha^{-1}\|y_0\|_{L^\infty(\Omega)}).$$

Hence

$$y(x,t) \leqq \|y_0\|_{L^\infty(\Omega)} - \alpha t \quad \forall x \in \Omega, \quad t \in (0, \alpha^{-1}\|y_0\|_{L^\infty(\Omega)})$$

and therefore

$$|y(x,t)| \leqq \|y_0\|_{L^\infty(\Omega)} - \alpha t \quad \forall (x,t) \in \Omega \times [0, \alpha^{-1}\|y_0\|_{L^\infty(\Omega)}].$$

Hence

$$(3.6)\qquad \phi(y_0) \leqq \alpha^{-1}\|y_0\|_{L^\infty(\Omega)} \quad \forall y_0 \in L^\infty(\Omega).$$

Let $S(t)$ be the semigroup generated on $L^2(\Omega)$ by the operator $-(A+F)$, i.e., $S(t)y_0 = y(t)$ where y is the solution to boundary value problem

$$\begin{aligned} &\frac{\partial y}{\partial t} - \Delta y + \beta(y) = 0 \quad \text{in } \Omega \times R^+,\\ &y(0) = y_0 \quad \text{in } \Omega \qquad y = 0 \quad \text{in } \partial\Omega \times R^+. \end{aligned}$$

If z is the solution to the linear parabolic equation

$$\begin{aligned} &\frac{\partial z}{\partial t} - \Delta z = 0 \quad \text{in } \Omega \times R^+,\\ &z(0) = |y_0| \quad \text{in } \Omega \qquad z = 0 \quad \text{in } \partial\Omega \times R^+, \end{aligned}$$

it follows by the maximum principle that

$$|y(x,t)|\leqq z(x,t) \quad \forall (x,t)\in \Omega\times R^+,$$

i.e.,

$$\|S(t)y_0\|_{L^\infty(\Omega)}\leqq Ct^{-N/2}\|y_0\|_{L^2(\Omega)} \quad \forall t>0 \tag{3.7}$$

because, as is well known,

$$\|S_0(t)y_0\|_{L^\infty(\Omega)}\leqq Ct^{-N/2}\|y_0\|_{L^2(\Omega)} \quad \forall t>0 \tag{3.8}$$

where C is a positive constant independent of y_0. (Here $S_0(t)$ is the C_0-semigroup generated by $-A$ on $L^2(\Omega)$.) Then taking into account the obvious inequality

$$\phi(y_0)\leqq \phi(S(t)y_0)+t \quad \forall t>0, \quad y_0\in L^2(\Omega), \tag{3.9}$$

we infer that $\phi(y_0)<\infty$ for all $y_0\in L^2(\Omega)$. (As a matter of fact we may extend ϕ on all of $L^1(\Omega)$.)

Now let $y_0\in L^2(\Omega)$ be arbitrary but fixed and let $\{y_0^n\}\subset L^2(\Omega)$ be such that $y_0^n\rightarrow y_0$ weakly in $L^2(\Omega)$ for $n\rightarrow\infty$. Since the function ϕ is obviously weakly lower semicontinuous on $L^2(\Omega)$ (this is an immediate consequence of the fact that a time-optimal controller exists and follows by standard arguments) it remains to show that

$$\limsup \phi(y_0^n)\leqq \phi(y_0). \tag{3.10}$$

Keeping in mind inequality (3.9) and the fact that for each $t>0$ the operator $S(t)$ is weakly-strongly continuous on $L^2(\Omega)$, without any loss of generality we may assume that

$$y_0^n\rightarrow y_0 \quad \text{strongly in } L^2(\Omega). \tag{3.11}$$

Let $u_0\in \mathcal{U}$ be such that $y(t_0,u_0)=0$ where $t_0=\phi(y_0)$ and let y_n be the corresponding solution to problem (1.3), i.e.,

$$\frac{\partial y_n}{\partial t}-\Delta y_n+\beta(y_n)=u_0 \quad \text{in } \Omega\times R^+,$$

$$y_n(0)=y_0^n \quad \text{in } \Omega, \qquad y_n=0 \quad \text{in } \partial\Omega\times R^+.$$

By (3.6), (3.7), and (3.9) we have

$$\begin{aligned}\phi(y_0^n)&\leqq \phi(y_0)+\phi(y_n(t_0))\\ &\leqq \phi(y_0)+\phi(S(\varepsilon)y_n(t_0))+\varepsilon\\ &\leqq \phi(y_0)+\varepsilon+C\varepsilon^{-N/2}\alpha^{-1}\|y_n(t_0)\|_{L^2(\Omega)}.\end{aligned} \tag{3.12}$$

On the other hand, it follows by (3.11) that $y^n(t)\rightarrow y(t)$ uniformly on $[0,t_0]$ in the strong topology of $L^2(\Omega)$ where $y(t)=y(t,y_0,u_0)$. Hence $\|y_n(t_0)\|_{L^2(\Omega)}\rightarrow 0$ as $n\rightarrow\infty$. If in (3.12) we take $\varepsilon=\varepsilon_n$ such that

$$\varepsilon_n\rightarrow 0, \quad \varepsilon_n^{-N/2}\|y_n(t_0)\|_{L^2(\Omega)}\rightarrow 0 \quad \text{for } n\rightarrow\infty$$

we get (3.10) as desired.

We note parenthetically that we have proved that the minimal time function ϕ for system (1.3) is weakly continuous on $L^2(\Omega)$ under the weaker assumption that β is a maximal monotone graph in $R\times R$ such that $0\in\beta(0)$.

Thus Theorem 1 is applicable in the present situation.

COROLLARY 1. *Under assumption* (3.2) *the dynamic programming equation of the time-optimal control problem for system* (3.1)

$$\alpha\|D\phi(y)\|_{L^1(\Omega)}+(Ay+Fy,D\phi(y))=1 \quad \textit{in } L^2(\Omega)\backslash\{0\} \tag{3.13}$$

has a unique strong viscosity solution satisfying conditions

$$\phi(0)=0, \quad \phi(y)\geqq 0 \quad \forall y\in L^2(\Omega), \tag{3.14}$$

namely, the minimal time function (3.3).

4. The time-optimal problem for boundary control of the heat equation. Consider the boundary control system

$$\begin{aligned} &\frac{\partial y}{\partial t}-\Delta y=0 \quad \text{in } \Omega\times(0,\infty),\\ &y(x,0)=y_0(x), \qquad x\in\Omega,\\ &y(x,t)=u(x,t) \quad \forall(x,t)\in\partial\Omega\times(0,\infty) \end{aligned} \tag{4.1}$$

where Ω is a bounded, open, connected set of R^n with C^∞-boundary, and $u\in L^\infty(\partial\Omega\times(0,\infty))$ are subject to constraints

$$|u(x,t)|\leqq\alpha \quad \text{a.e. } (x,t)\in\partial\Omega\times(0,\infty). \tag{4.2}$$

The maximum principle for the time-optimal problem associated with boundary control system (4.1), (4.2) has been obtained by Fattorini [13].

If we denote by $A: L^2(\Omega)\to L^2(\Omega)$ the operator $-\Delta$ with homogeneous Dirichlet conditions, i.e., $D(A)=H_0^1(\Omega)\cap H^2(\Omega)$ and for $u\in L^2(\Omega)$ we set $\theta=Bu$, where $\theta\in L^2(\Omega)$ is the weak solution to Dirichlet problem

$$\Delta\theta=0 \quad \text{in } \Omega, \qquad \theta=u \quad \text{in } \partial\Omega,$$

we may write system (4.1) as

$$\begin{aligned} &y'(t)+A(y(t)-Bu(t))=0 \quad \text{in } R^+,\\ &y(0)=y_0. \end{aligned}$$

Equivalently,

$$\begin{aligned} &z'(t)+Az(t)=Bu(t) \quad \text{in } R^+,\\ &z(0)=A^{-1}y_0=z_0 \end{aligned} \tag{4.3}$$

where $z=A^{-1}y$.

Let $\phi: L^2(\Omega)\to R^+$ be the minimal time function corresponding to system (4.3), i.e.,

$$\begin{aligned} \phi(z_0)=\inf\{T;\ \exists u\in L^\infty(\partial\Omega\times(0,\infty)), &|u|\leqq\alpha\\ &\text{a.e. in } \partial\Omega\times(0,\infty),\ z(T)=0\}. \end{aligned} \tag{4.4}$$

Let us first show that ϕ is everywhere finite and weakly continuous on $L^2(\Omega)$. Indeed by the results of Russell [16], $\phi(z_0)<\infty$ for all $z_0\in D(A)$.

Then, by the obvious inequality

$$\phi(z_0)\leqq\phi(e^{-A\varepsilon}z_0)+\varepsilon \quad \forall z_0\in L^2(\Omega), \tag{4.5}$$

we infer that $\phi<\infty$ on $L^2(\Omega)$, as claimed. Applying Theorem 1 in [7], it follows that the minimal time function ϕ is continuous on $L^2(\Omega)$, and since the semigroup e^{-At} is compact for $t>0$ we conclude that ϕ is weakly continuous on $L^2(\Omega)$.

Recalling that B^* (the adjoint operator of B) is given by $B^*p=-(\partial/\partial\nu)A^{-1}p$ in $\partial\Omega$ for $p\in L^2(\Omega)$, we may write the Bellman equation associated with system (4.3)

$$\alpha\left\|\frac{\partial}{\partial\nu}A^{-1}D\phi(y)\right\|_{L^1(\partial\Omega)}+(Ay, D\phi(y))=1 \quad \text{for } y\neq 0. \tag{4.6}$$

Applying Theorem 1, we find Corollary 2.

COROLLARY 2. *The minimal time function ϕ is the unique weakly continuous strong viscosity solution to* (4.6) *in $L^2(\Omega)\backslash\{0\}$ satisfying the conditions*

$$\phi(0)=0,\quad \phi\geqq 0 \quad \textit{in } L^2(\Omega). \tag{4.7}$$

Now if $\psi(y)=\phi(A^{-1}y)$ is the minimal time function associated with system (4.1) we see by (4.6) that, formally, it satisfies the equation

$$\alpha\left\|\frac{\partial}{\partial\nu}D\psi(y)\right\|_{L^1(\Omega)}+(Ay, D\psi(y))=1. \tag{4.8}$$

Then Corollary 2 implies the existence and uniqueness of a viscosity solution $\psi:(D(A))^*\to(0,\infty)$ for (4.8) in $(D(A))^*\backslash\{0\}$. (We have denoted by $(D(A))^*$ the dual space of $D(A)$.)

Acknowledgment. I am indebted to Professor P. L. Lions for some useful remarks on an earlier version of this paper.

REFERENCES

[1] V. BARBU, *Nonlinear Semigroups and Differential Equations in Banach Spaces*, Noordhoff, Leyden, 1976.

[2] ———, *The time optimal problem for a class of nonlinear distributed systems*, Lasiecka and Triggiani, eds., Lecture Notes in Control and Information Theory, Vol. 97, Springer-Verlag, Berlin, New York, 1987, pp. 15-48.

[3] V. BARBU AND T. PRECUPANU, *Convexity and Optimization in Banach Spaces*, D. Reidel, Dordrecht, 1986.

[4] M. BARDI, *Boundary value problem for the minimum-time function*, SIAM J. Control Optim., 4 (1989), pp. 776-785.

[5] H. BRÉZIS, *Opérateurs maximaux monotones et semigroupes de contractions dans les espaces de Hilbert*, Mathematical Studies, Vol. 5, North-Holland, Amsterdam, 1973.

[6] O. CARJĂ, *On the minimal time function for distributed control system in Banach spaces*, J. Optim. Theory Appl., 44, (1984), pp. 397-406.

[7] ———, *On continuity of the minimal time function for distributed control systems*, Boll. Un. Mat. Ital. (6), 6 (1985), pp. 293-302.

[8] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.

[9] ———, *Hamilton-Jacobi equations in infinite dimensions*. I. *Uniqueness of viscosity solutions*, J. Funct. Anal., 62 (1985), pp. 379-396.

[10] ———, *Solutions de viscosité pour les équations de Hamilton-Jacobi en dimension infinie intervenant dans le contrôle optimal de problèmes d'évolution*, C. R. Acad. Sci. Paris, 305 (1987), pp. 233-236.

[11] ———, *Solutions of Hamilton-Jacobi equations in infinite dimensions. Part.* IV. *Hamiltonian with unbounded linear terms*, J. Funct. Anal., 87 (1989).

[12] M. G. CRANDALL, H. ISHII, AND P. L. LIONS, *Uniqueness of viscosity solutions of Hamilton-Jacobi equations, revised*, J. Math. Soc. Japan, 39 4 (1987), pp. 581-596.

[13] H. D. FATTORINI, *The time-optimal problem for boundary control of the heat equation*, in Calculus of Variations and Control Theory, Academic Press, New York, San Francisco, London, 1976, pp. 305-320.

[14] S. N. KRUŽKOV, *Generalized solutions of the Hamilton-Jacobi equations of eikonal type*, Math. USSR-Sb., 27 (1975), pp. 406-446.

[15] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Research Notes in Mathematics, Vol. 69, Pitman, Boston, 1982.

[16] D. L. RUSSELL, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., 52 (1973), pp. 189-211.

[17] D. TĂTARU, *Viscosity solutions of Hamilton-Jacobi equations with unbounded nonlinear terms*, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 457-473, March 1991

EXISTENCE OF CONTROL LYAPUNOV FUNCTIONS AND APPLICATIONS TO STATE FEEDBACK STABILIZABILITY OF NONLINEAR SYSTEMS*

J. TSINIAS†

Abstract. The asymptotic and practical stabilization for the affine in the control nonlinear systems, which extends the results of Artstein, Sontag, and Tsinias is explored. Sufficient conditions for the existence of control Lyapunov functions are presented guaranteeing stabilization. The corresponding feedback laws are smooth, except possibly at the equilibrium of the system.

Key words. control Lyapunov functions, state feedback, stabilizability.

AMS(MOS) subject classification. 93D15

1. Introduction.

Feedback stabilization of nonlinear systems is a problem of great importance in control theory. Geometric, decomposition, and linearization methods, Lyapunov and center manifold theorems, as well as sliding mode techniques have been used by many authors to derive necessary and sufficient conditions for stabilization of systems at a specified equilibrium. Regularity assumptions of the stabilizing feedback near the equilibrium play an important role in the theory that has been developed (see, for instance, [1]-[14], [16]-[19], [22], [23], [27]-[42]). Linear, smooth, almost smooth, sliding mode, and piecewise-analytic controllers have been used in the previously mentioned works where the various types of regularity requirements lead to many different notions of stabilization. In particular, in [3], [29], [31], [32], [37]-[39] the corresponding necessary and sufficient conditions for stabilization are of Lyapunov type, the proposed stabilizers are smooth except possibly at the equilibrium, and their construction is based on the existence of an appropriate control Lyapunov function.

Our main purpose is to explore further the asymptotic and practical stabilization problem for a wide class of affine in the control nonlinear systems. This class can be characterized in terms of computable control Lyapunov functions, which in certain cases depend directly on the dynamics of the system. The main techniques used to do this are Lyapunov's direct methods [15], [21], [26], [45] and the converse stability theorems of Massera [24], Kurzweil [20], and Wilson [43], [44]. It will be useful to recall here the precise definitions of the notions of stabilization, control Lyapunov functions, and the most important results provided in [3], [31], [37], and [38], as well as some further generalizations.

Consider the system

$$\dot{x}=f(x)+\sum_{i=1}^{l} u_i g_i(x)=f(x)+g(x)u, \tag{1.1}$$

where the state space is R^n and the control $u=(u_1,\cdots,u_l)'$ takes values on R^l. We assume that $0\in R^n$ is an equilibrium for f. The vector fields $f, g_1,\cdots,g_l$ are supposed to be smooth (C^∞).

We say that (1.1) is *asymptotically stabilizable* (at zero) by means of the feedback law $u=k(x)$, if $0\in R^n$ is an asymptotically stable equilibrium for the resulting closed-loop system

$$\dot{x}=f(x)+g(x)k(x). \tag{1.2}$$

* Received by the editors November 20, 1989; accepted for publication (in revised form) May 9, 1990.

† Department of Mathematics, National Technical University, Zografou Campus, 15773 Athens, Greece.

System (1.1) is said to be *practically stabilizable* (at zero) by means of the family of feedback laws $\{k_r, r>0\}$, if for any sufficiently small r and for any x_0 near the open sphere $S(0, r)$ of radius r around zero, the corresponding trajectory $x_r(t, x_0)$ of the closed-loop system

$$\dot{x} = f(x) + g(x)k_r(x) \tag{1.3}$$

enters $S(0, r)$ after some time $t = T < +\infty$ and it stays in this region thereafter. We note that the notion of practical stability is discussed by La Salle and Lefschetz [21].

DEFINITION 1.1. The system (1.1) satisfies the *Lyapunov condition* at zero (L.C.), if there exist a neighborhood N of $0 \in R^n$ and a real function $\Phi: N \to R$, which is at least continuously differentiable on N, is positive definite, i.e., $\Phi(0) = 0$ and $\Phi(x) > 0$ for $x \in N - \{0\}$, and such that for any $x \in N - \{0\}$ the following condition holds:

$$g_i(\Phi)(x) \overset{\text{def}}{=} D\Phi g_i(x) = 0, \qquad i = 1, \cdots, l \Rightarrow f(\Phi)(x) < 0, \tag{1.4}$$

where $D\Phi$ denotes the derivative of Φ. A continuously differentiable real function Φ is called a *control Lyapunov function*, if it is positive definite and satisfies condition (1.4).

We say that the control Lyapunov function Φ above satisfies the *bounded* control property, if there exists a positive real function $d: N \to R$ such that d is bounded on N and for every $x \in N - \{0\}$ there exists a vector $u \in R^l$ satisfying the following inequalities:

$$\|u\| < d(x), \tag{1.5}$$

$$f(\Phi)(x) + g(\Phi)(x)u < 0, \tag{1.6}$$

where $g(\Phi) \overset{\text{def}}{=} (g_1(\Phi), \cdots, g_l(\Phi))$ and $\| \ \|$ is the usual Euclidean norm. If, in addition, $d(x) \to 0$ as $x \to 0$, then we say that Φ satisfies the *small* control property.

The next theorem was originally established in [3]. Versions and generalizations of this theorem can also be found in [31], [37]-[39].

THEOREM 1.2. (i) *The system* (1.1) *satisfies the* L.C., *if and only if it is asymptotically stabilizable at zero by means of a feedback law $u = k(x)$, which is smooth for $x \neq 0$ near zero.*

(ii) *The corresponding control Lyapunov function Φ satisfies the bounded control property, if and only if there exists a stabilizer $u = k(x)$, which is smooth for $x \neq 0$ near zero and satisfies $\|k(x)\| < d(x)$, where d is defined in* (1.5). *It turns out that k is bounded in a neighborhood of zero and if, in addition, Φ satisfies the small control property then $k(x) \to 0$ as $x \to 0$.*

We note that the smoothness property of the stabilizing feedback laws constructed in [31] and [37] requires the smoothness of the Lyapunov function Φ, since they depend directly on the Lie derivatives $f(\Phi)$ and $g(\Phi)$. In [3] the regularity properties of the proposed feedback laws require only the differentiability of Φ.

DEFINITION 1.3. The system (1.1) satisfies the *practical Lyapunov condition* (P.L.C.), if there exist a family of continuously differentiable mappings

$$\Phi_r: R^n \to R, \quad \Phi_r(0) = 0, \quad r > 0,$$

and compact neighborhoods N and N_r of $0 \in R^n$ such that $N_r \subset S(0, r) \subset N_r$, Φ_r is strictly positive for x outside N_r and for any sufficiently small r condition (1.4) holds with $\Phi = \Phi_r$ and $x \in N_r - N_r$ and further

$$\min\{\Phi_r(x), x \in \partial N_r\} > \max\{\Phi_r(x), x \in S[0, r]\}, \tag{1.7a}$$

where ∂N is the boundary of N. Moreover, for any sequences $\{r_i\} \subset R^+$ and $\{x_{r_i}\} \subset N$ with $\lim r_i = 0$ and $\lim x_{r_i} = a$ as $i \to +\infty$, we have

$$\overline{\lim}\, \Phi_{r_i}(x_{r_i}) = 0 \quad \text{for } a = 0, \quad \text{and} \quad \underline{\lim}\, \Phi_{r_i}(x_{r_i}) > 0 \quad \text{for } a > 0. \tag{1.7b}$$

The corresponding family $\{\Phi_r\}$ of the mappings above is called *control Lyapunov family.*

We say that the control Lyapunov family $\{\Phi_r\}$ satisfies the *bounded* control property, if there exist a positive real function $d(x)$, which is bounded on a neighborhood N of zero, and a family of positive real numbers $\{c_r, r>0\}$ such that $c_r \to 0$ as $r \to 0$ and for any sufficiently small r conditions (1.6) and (1.7) hold with $\Phi = \Phi_r$, $x \in N_{2r} - N_{1r}$ and for some vector $u \in R^l$ with

$$\|u\| < d(x) + c_r.$$

If, in addition, $d(x) \to 0$ as $x \to 0$, then we say that $\{\Phi_r\}$ satisfies the *small* control property.

The following theorem presents sufficient conditions for practical stabilization. Its proof follows by using arguments similar for those of the proof of Theorem 1.2.

THEOREM 1.4. (i) *If the system* (1.1) *satisfies the* P.L.C., *then it is practically stabilizable at zero by means of a family of smooth feedback laws* $\{k_r(x)\}$.

(ii) *Moreover, if the control Lyapunov family* $\{\Phi_r\}$ *satisfies the bounded* (*small*) *control property, then for any r the corresponding stabilizer* k_r *satisfies the inequality*

$$\|k_r(x)\| < d(x) + c_r, \tag{1.8}$$

for x near zero, where d and c_r *are as given in Definition* 1.3.

Obviously, if (1.1) satisfies the L.C., it also satisfies the P.L.C. Furthermore, as a consequence of Proposition 4 of [38] and the corollary in paragraph 7 of Sontag [32] we obtain the following result.

PROPOSITION 1.5. *If the system* (1.1) *is asymptotically stabilizable by means of the feedback* $u = k(x)$, *which is continuous in a neighborhood of zero, then it is practically stabilizable by means of a family of smooth feedback laws.*

In § 2 we consider systems (1.1) of the form

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix} + \begin{pmatrix} 0 \\ g_2(x_1, x_2) \end{pmatrix} u, \tag{1.9}$$
$$x = (x_1', x_2')' \in R^{n_1} \times R^{n_2}, \quad n_1 + n_2 = n, \quad u \in R^l$$

and we derive sufficient conditions for the existence of control Lyapunov functions guaranteeing stabilization. We note that similar decompositions have been considered by Vidyasagar and other authors [41], [42], [36], [10] but their methodology is quite different. Let us explain briefly the main idea of § 2 by the following interesting case. Consider a single input system (1.9) with $n_2 = 1$, namely, g_2 is a real function. Let us assume that there is a continuously differentiable map $v = \phi(x_1)$ with $\phi(0) = 0$, which is the unique solution of the equation $g_2(x_1, v) = 0$ and simultaneously asymptotically stabilizes the system $\dot{x}_1 = f_1(x_1, v)$, $v \in R^{n_2}$ at zero (see Example 4.3 of this paper). Then the system (1.9) satisfies the L.C. and so by Theorem 1.2 it is asymptotically stabilizable at $0 \in R^n$. Indeed, consider a smooth Lyapunov function $V(x_1)$ for the closed-loop system $\dot{x}_1 = f_1(x_1, \phi(x_1))$ and define

$$\Phi(x) = V(x_1) + \tfrac{1}{2}(x_2 - \phi(x_1))^2.$$

Obviously, Φ is continuously differentiable and positive definite. Moreover, Φ is a control Lyapunov function for (1.9). Indeed, for any $x \neq 0$ with

$$g_2(\Phi)(x) = (x_2 - \phi(x_1)) g_2(x_1, x_2) = 0$$

it follows that $x_2 = \phi(x_1)$; therefore,

$$\begin{aligned} f(\Phi)(x) &= f_1(V)(x_1, x_2) + (x_2 - \phi(x_1)) f_2(x_1, x_2)(x_2 - \phi(x_1)) D\phi(x_1) f_1(x_1, x_2)|_{x_2 = \phi(x_1)} \\ &= f_1(V)(x_1, \phi(x_1)) < 0. \end{aligned}$$

It turns out that (1.9) satisfies the L.C. and so by Theorem 1.2 it is stabilizable at zero by meaŋs of a feedback law which is smooth for $x \neq 0$ near zero. The previous discussion is applicable to systems of the form

$$\overset{(n)}{x} = f_2(x, \overset{(1)}{x}, \cdots, \overset{(n-1)}{x}) + ug_2(x, \overset{(1)}{x}, \cdots, \overset{(n-1)}{x}), \quad x \in R \quad \text{where } \overset{(i)}{x} = \frac{d^i x}{dt^i},$$

or equivalently

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_n \end{pmatrix} = \begin{pmatrix} x_2 \\ x_3 \\ \vdots \\ f_2(x) \end{pmatrix} + u \begin{pmatrix} 0 \\ 0 \\ \vdots \\ g_2(x) \end{pmatrix}, \qquad f_2(0) = g_2(0) = 0.$$

The above system satisfies the L.C. provided that $x_n = \phi(x_1, x_2, \cdots, x_{n-1})$, $\phi(0) = 0$ is the unique solution of $g_2(x) = 0$ and $0 \in R^{n-1}$ is an asymptotically stable equilibrium for the lower-dimensional system

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \\ \vdots \\ \dot{x}_{n-1} \end{pmatrix} = \begin{pmatrix} x_2 \\ x_3 \\ \vdots \\ \phi(x_1, x_2, \cdots, x_{n-1}) \end{pmatrix}.$$

In § 2 we extend the above analysis to more general cases. In particular, in Theorems 2.1 and 2.2 we provide sufficient conditions for the existence of a control Lyapunov function depending only on the stability behavior of the mappings f_1 and g_2.

Special emphasis is given to asymptotic and practical stabilization for systems (1.9), whose control term g_2 is constant and have the form

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix} + \begin{pmatrix} 0 \\ u \end{pmatrix}, \qquad (x_1, x_2) \in R^{n_1} \times R^{n_2}, \quad u \in R^{n_2}. \tag{1.10}$$

Proposition 2.3 and Theorems 2.4 and 2.6 of this paper generalize some well-known results of [23], [34], and [37] concerning the case (1.10).

Section 3 deals with the relationship between the Lyapunov conditions and the position on the complex plane of the uncontrollable eigenvalues of the linearization of (1.1) at zero. Finally, in § 4 we illustrate the theory we develop by four numerical examples.

2. Main results.

To begin with, we state and prove a theorem, where the L.C. is characterized in terms of suitable positive functions concerning the lower-dimensional subsystems of (1.9):

$$\dot{x}_1 = f_1(x_1, v), \qquad v \in R^{n_2}, \tag{2.1}$$

$$\dot{x}_2 = g_2(x_1, x_2)u, \qquad u \in R^l. \tag{2.2}$$

THEOREM 2.1. (i) *Suppose that there exist neighborhoods N and N_1 of $0 \in R^n$ and $0 \in R^{n_1}$, respectively, and mappings $r: N \to R^l$, $\phi: N_1 \to R^{n_2}$, and $W: N \to R$ such that $\phi(0) = 0$, ϕ is continuous, and W is continuously differentiable, and let us denote*

$$M = \{x \in N\colon x_2 = \phi(x_1)\}, \qquad S_1 = \{x \in N\colon W(x) = 0\}, \quad \textit{and}$$

$$S_2 = \left\{x \in N\colon \left(\frac{\partial W}{\partial x_2} g_2 r\right)(x) = 0\right\}.$$

Assume that

(a) *For any* $x=(x_1', x_2')'$ *near zero it follows that*

$$W(x) \geqq 0 \quad \text{and} \quad S_2 \subset S_1 = M.$$

(b) *The system* (2.1) *is asymptotically stabilizable at* $0 \in R^{n_1}$ *by means of the feedback law* $v=\phi(x_1)$.

Under the previous assumptions (a) *and* (b) *the system* (1.9) *satisfies the* L.C. *Let* Φ *be the corresponding control Lyapunov function.*

(ii) *Let us further assume that r is bounded,* ϕ *and* $\partial W/\partial x_2$ *continuously differentiable, and*

(c) *There exist positive constants c and* $a \leqq 2$ *such that*

$$\left(\frac{\partial W}{\partial x_2} g_2 r\right)(x) \leqq -c\|x_2-\phi(x_1)\|^a, \tag{2.3}$$

for x near $0 \in R^n$.

Then, if $a \leqq 1$ *the function* Φ *satisfies the small control property, and if* $1 < a \leqq 2$ *and in addition* $0 \in R^{n_1}$ *is an exponentially stable equilibrium for the closed-loop system* $\dot{x}_1 = f_1(x_1, \phi(x_1))$, *then the control Lyapunov function* Φ *satisfies the bounded control property.*

Proof. (i) Since $0 \in R^{n_1}$ is an asymptotically stable equilibrium for the system

$$\dot{x}_1 = f_1(x_1, \phi(x_1)) \tag{2.4}$$

then the converse Lyapunov theorem of Kurzweil [20] asserts that there exists a smooth Lyapunov function $V: R^{n_1} \to R$ such that $V(0)=0$, $V(x_1)>0$ and

$$DVf_1(x_1, \phi(x_1)) < 0$$

for $x_1 \neq 0$ near zero. Obviously, by assumption (a) the function

$$\Phi = V + W$$

is positive definite. Let $f=(f_1', f_2')'$ and $g=(0', g_2')'$. Then

$$f(\Phi)=f_1(V)+f(W) \quad \text{and} \quad g(\Phi)=\frac{\partial W}{\partial x_2} g_2.$$

For any $x \neq 0$ such that $g(\Phi)(x)=((\partial W/\partial x_2)g_2)(x)=0$ it follows by assumption (a) that $x_2=\phi(x_1)$; hence,

$$DW(x_1, \phi(x_1))=0$$

and so

$$f(\Phi)(x)=DVf_1(x_1, \phi(x_1))<0.$$

Therefore (1.9) satisfies the L.C. and $\Phi = V+W$ is a control Lyapunov function.

(ii) Let us assume that (2.3) holds and let $a \leqq 1$. Define

$$q(x) = -DVf_1(x_1, \phi(x_1)) + \|x_2-\phi(x_1)\|^2.$$

Clearly, q is positive definite. Since f and $\partial W/\partial x_2$ are continuously differentiable and $DW(x_1, \phi(x_1))=0$, there are positive contants c_1 and c_2 such that

$$\|DW(x)\| \leqq c_1\|x_2-\phi(x_1)\|, \tag{2.5}$$

$$\|Df(x)\| \leqq c_2 \quad \text{for } x \text{ near zero.} \tag{2.6}$$

Let $b(x) = c_2\|DV(x_1)\| + c_1\|f(x)\| + \|x_2 - \phi(x_1)\|$. Then by (2.3), (2.5), and (2.6) it follows that

$$\begin{aligned}|(f(\Phi)+q)(x)| &\leqq c_2\|DV(x_1)\|\,\|x_2-\phi(x_1)\| + c_1\|f(x)\|\,\|x_2-\phi(x_1)\| + \|x_2-\phi(x_1)\|^2\\ &\leqq b(x)\|x_2-\phi(x_1)\| \leqq b(x)\|x_2-\phi(x_1)\|^a \leqq -\frac{b(x)}{c}(D\Phi gr)(x)\end{aligned}$$

and, therefore,

$$\left\{f(\Phi)+g(\Phi)\left(\frac{rb}{c}\right)\right\}(x) \leqq -q(x) < 0 \quad \text{for } x \neq 0 \text{ near } \textit{zero}.$$

Since r is bounded and b is continuous with $b(0)=0$, it follows that Φ satisfies the small control property.

Let us finally assume that (2.3) holds with $1 < a \leqq 2$, ϕ is continuously differentiable, and $0 \in R^{n_1}$ is exponentially stable for (2.4). Since $\phi(0)=0$, there is a constant $c_3 > 0$ such that

$$\|\phi(x_1)\| < c_3\|x_1\|, \quad x_1 \text{ near } 0 \in R^{n_1} \tag{2.7}$$

and since $f(0)=0$ it follows by (2.6) and (2.7) that

$$\begin{aligned}\|f(x)\| &\leqq c_2(\|x_1\|+\|x_2\|) \leqq c_2(\|x_1\|+\|\phi(x_1)\|+\|x_2-\phi(x_1)\|)\\ &\leqq c_2((c_3+1)\|x_1\|+\|x_2-\phi(x_1)\|),\end{aligned} \tag{2.8}$$

for all x in a neighborhood of $0 \in R^n$. Exponential stability of $0 \in R^{n_1}$ with respect to (2.4) implies the existence of a continuously differentiable function V and a positive constant $c_4 > 0$ such that

$$DVf_1(x_1, \phi(x_1)) \leqq -c_4\|x_1\|^2, \quad \text{and} \tag{2.9}$$

$$\|DV(x_1)\| \leqq \|x_1\| \tag{2.10}$$

for x_1 near zero [15], [45]. Let

$$\begin{aligned}q(x) = -f_1(V)(x_1, \phi(x_1)) - DV\frac{\partial f_1}{\partial x_2}(x_1, \phi(x_1))(x_2-\phi(x_1))\\ -f(W)(x) + k\|x_2-\phi(x_1)\|^2, \qquad k > 0.\end{aligned}$$

Then (2.5)-(2.10) imply that there are constants M_1, $M_2 > 0$ such that

$$q(x) \geqq c_4\|x_1\|^2 - M_1\|x_1\|\,\|x_2-\phi(x_1)\| + (k-M_2)\|x_2-\phi(x_1)\|^2$$

and so q is positive definite, provided that k is a suitably large positive constant. Finally, let $L > 0$ such that

$$\left\|f(x)-f(x_1,\phi(x_1)) - \frac{\partial f_1}{\partial x_2}(x_1,\phi(x_1))(x_2-\phi(x_1))\right\| < L\|x_2-\phi(x_1)\|^2 \tag{2.11}$$

locally around $0 \in R^n$. From (2.11) and assumption (c), there is a positive constant $s > 0$ satisfying

$$\begin{aligned}|(f(\Phi)+q)(x)| = |f_1(V)(x) + f(W)(x) - f_1(V)(x_1, \phi(x_1))\\ -DV\frac{\partial f_1}{\partial x_2}(x_1,\phi(x_1))(x_2-\phi(x_1)) - f(W)(x) + k\|x_2-\phi(x_1)\|^2|\\ \leqq s\|x_2-\phi(x_1)\|^a \leqq -\frac{s}{c}(g(\Phi)r)(x)\end{aligned}$$

and so

$$\left\{f(\Phi)+g(\Phi)\left(\frac{rs}{c}\right)\right\}(x) \leqq -q(x)<0$$

for $x \neq 0$ locally around zero. Since rs/c is bounded, Φ satisfies the bounded control property. □

The next theorem is one of our main results and is a consequence of Theorem 2.1(i). It provides a control Lyapunov function that depends directly on the dynamics of the systems (2.1) and (2.2) and its proof is based on the converse Lyapunov theorem established by Wilson [43], [44].

THEOREM 2.2. *Suppose that there exist neighborhoods N_1 and N of $0 \in R^{n_1}$ and $0 \in R^n$, respectively, and mappings $r: N \to R^l$ and $\phi: N_1 \to R^{n_2}$ such that r is Lipschitz continuous, ϕ is continuous with $\phi(0)=0$, the origin $0 \in R^{n_1}$ is an asymptotically stable equilibrium for* (2.4) *and the set*

$$M=\{x \in N: x_2=\phi(x_1), x_1 \in N_1\}$$

is asymptotically stable with respect to

$$\dot{x}=(\hat{g})(x) \overset{\text{def}}{=} \begin{pmatrix} 0 \\ (g_2 r)(x) \end{pmatrix}, \qquad x=(x_1', x_2') \in R^{n_1} \times R^{n_2}. \tag{2.12}$$

Then (1.9) *satisfies the* L.C.

Proof. Consider the system (2.12) evolving on $X \overset{\text{def}}{=} \{x \in R^n: x_1 \in N_1\}$. Note that the restriction of (2.12) to the space R^{n_1} gives $\dot{x}_1=0$ and so the region X is positively invariant. According to [43] and [44], since $M \subset X$ is asymptotically stable with respect to (2.12), there exists a smooth Lyapunov function $W: X \to R$ of M, namely, $W(x)>0$ and $(DW\hat{g})(x)<0$ for $x \notin M$, whereas $W(x)=0$ for $x \in M$. It turns out that $((\partial W/\partial x_2) g_2 r)(x)<0$ for $x_2 \neq \phi(x_1)$ near zero and $((\partial W/\partial x_2) g_2 r)(x)=0$ if and only if $x_2=\phi(x_1)$. Therefore, the assumptions (a) and (b) of Theorem 2.1 are fulfilled and the system (1.9) satisfies the L.C. □

Next we specialize Theorem 2.1 to the particular case of systems (1.10). In this case without loss of generality we assume that $f_2 \equiv 0$. Indeed, otherwise we apply the smooth law $u \to -f_2+u$ and the system (1.10) becomes

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix}=\begin{pmatrix} f_1(x_1, x_2) \\ 0 \end{pmatrix}+\begin{pmatrix} 0 \\ I \end{pmatrix} u. \tag{2.13}$$

It is well known (see, for instance, [23], [34], [37]) that if (2.1) is asymptotically stabilizable by means of the smooth feedback

$$v=\phi(x_1), \qquad \phi(0)=0,$$

then (2.13) is also smoothly asymptotic stabilizable (hence there exists a control Lyapunov function satisfying the small control property). If the smoothness requirements for the map ϕ are relaxed, further generalizations are possible. For instance, let us assume that ϕ is continuously differentiable. Then, as in Theorems 3 and 4 of [37], it can be shown that (2.13) admits a control Lyapunov function satisfying the small control property. The following proposition asserts that the last statement is also an immediate consequence of Theorem 2.1(ii).

PROPOSITION 2.3. *Consider the system* (2.13) *and suppose that* (2.1) *is asymptotically stabilizable by means of the feedback law $v=\phi(x_1)$, $\phi(0)=0$, which is continuously differentiable in a neighborhood of $0 \in R^{n_1}$. Then* (2.13) *satisfies the* L.C. *and the corresponding control Lyapunov function satisfies the small control property.*

Proof. Let V be a smooth Lyapunov function for (2.4) and let

$$W(x)=\tfrac{1}{2}\|x_2-\phi(x_1)\|^2.$$

Obviously, W is nonnegative definite and $\partial W/\partial x_2$ continuously differentiable near zero. Define $r=(r_1,\cdots,r_{n_2})'$ with $r_i(x)=-\mathrm{sgn}\,(x_2-\phi(x_1))_i$. Then

$$\left(\frac{\partial W}{\partial x_2}g_2 r\right)(x)=\left(\frac{\partial W}{\partial x_2}r\right)(x)\leqq -\|x_2-\phi(x_1)\|,$$

r is uniformly bounded on R^n and so W satisfies the properties (a) and (b) of Theorem 2.1. In particular, (2.3) holds with $a=1$; therefore, the function $\Phi=V+W$ satisfies the small control property. □

In the rest part of the paper we deal with the practical stabilization problem for the system (2.13). We shall establish that under the only assumption that the stabilization law $v=\phi$ for (2.1) is *Lipschitz continuous*, then there exists a control Lyapunov family, which satisfies the small control property of Definition 1.3. Therefore, in that case practical stabilization is possible by means of a family of feedback laws $\{k_r\}$ satisfying (1.8).

THEOREM 2.4. *Let us assume that* (2.1) *is asymptotically stabilizable by means of the feedback law* $v=\phi(x_1)$ *with* $\phi(0)=0$. *Then*

(i) *If* ϕ *is continuous near zero, then* (2.13) *satisfies the* P.L.C.

(ii) *If* ϕ *is Lipschitz continuous, then* (2.13) *satisfies the* P.L.C. *and, further, the corresponding control Lyapunov family satisfies the small control property.*

In order to prove statement (ii) of Theorem 2.4 we need the following lemma.

LEMMA 2.5. *Let* $\phi: R^n\to R^m$ *be a real map that satisfies a Lipschitz condition on a compact neighborhood* U *of* $0\in R^n$, *and* $\phi(0)=0$. *Then there exist a neighborhood* Q *of* $0\in R^n$, *a constant* $K>0$ *and a family of smooth mappings* $\phi_\varepsilon: Q\to R^m$, $\varepsilon>0$ *such that for every sufficiently small* ε *the following hold*:

$$\|\phi(x)-\phi_\varepsilon(x)\|<\varepsilon \quad \forall x\in Q-S(0,\varepsilon), \tag{2.14}$$

$$\phi_\varepsilon(0)=0, \tag{2.15}$$

$$\|D\phi_\varepsilon(x)\|<K \quad \forall x\in Q. \tag{2.16}$$

Moreover, for any sequences $\{\varepsilon_i\}\subset R^+$ *and* $\{x_i\}\subset Q$ *such that* $\varepsilon_i\to 0$ *and* $x_i\to x\in Q$, *as* $i\to+\infty$, *there exist subsequences* $\{\varepsilon_i'\}$ *and* $\{x_i'\}$ *of* $\{\varepsilon_i\}$ *and* $\{x_i\}$, *respectively, and a positive constant* $0\leqq\sigma\leqq 1$ *satisfying*

$$\phi_{\varepsilon_i'}(x_1')\to\sigma\phi(x). \tag{2.17}$$

Proof. Let Q be a compact neighborhood of zero, which is contained in the interior of U and let $p: R^n\to R$ be a smooth real map such that $p(x)=1$ for $x\in Q$ and $p(x)=0$ outside U. We define $\hat{\phi}=p\phi$ and consider its Sobolev's regularization

$$h_s(x)\overset{\text{def}}{=}\int a_s(x-u)\hat{\phi}(u)\,du=\int_{\|u\|<s} a_s(u)\hat{\phi}(x-u)\,du, \qquad 0<s<1,$$

where $a_s(x)=s^{-n}a(s^{-1}\|x\|)$, a being any positive smooth mapping such that $a(\|x\|)\leqq 1$, $a(\|x\|)=0$ for $\|x\|\geqq 1$ and $\int a(\|x\|)\,dx=1$. Using standard arguments (see, for instance, [25]), we can easily establish that, since $\hat{\phi}$ is continuous and has bounded support, the sequence h_s converges to $\hat{\phi}=p\phi=\phi$ as $s\to 0$ uniformly with respect to $x\in Q$, and further h_s is smooth for every s. Since p is smooth having bounded support in U and

ϕ is Lipschitz continuous on Q, it follows that $\hat{\phi}=p\phi$ also satisfies a Lipschitz condition in $S(Q,1)$, namely, there is a positive constant c such that

$$\|\hat{\phi}(v_1)-\hat{\phi}(v_2)\|<\frac{c}{2}\|v_1-v_2\| \quad \forall v_1, v_2\in S(Q,1).$$

For each $x_1, x_2\in Q$, $1>s>0$, and $\|u\|<s$ we have x_1-u, $x_2-u\in S(Q,1)$ and so

$$\|h_s(x_1)-h_s(x_2)\|\leqq\int_{\|u\|<s} a_s(u)\|\hat{\phi}(x_1-u)-\hat{\phi}(x_2-u)\|\,du$$

$$\leqq\frac{c}{2}\|x_1-x_2\|\int_{\|u\|<s} a_s(u)\,du=\frac{c}{2}\|x_1-x_2\|.$$

Smoothness of h_s and the previous inequality imply that

$$\|Dh_s(x)\|<c, \quad \text{and}$$

$$\|h_s(x)\|\leqq c\|x\|+\|h_s(0)\| \quad \forall x\in Q \text{ and } 0<s<1.$$

Consider, finally, a smooth real function ψ satisfying $\psi(x)=1$ for $\|x\|>1$, $\psi(x)=0$ for $\|x\|<\frac{1}{2}$, and $\psi(x)\leqq 1$ otherwise. Let $L>0$ such that $\|D\psi(x)\|<L$ for every $x\in R^n$, and let $s=s(\varepsilon)\leqq\varepsilon$ satisfying $\|\phi(x)-h_s(x)\|<\varepsilon$ for all $x\in Q$. We define

$$\phi_\varepsilon(x)=\psi\left(\frac{x}{\varepsilon}\right)h_{s(\varepsilon)}(x)$$

and $K=Lc+L+c$. Obviously $\phi_\varepsilon(x)=h_s(x)$ for each $x\in Q-S(0,\varepsilon)$ and $\phi_\varepsilon(0)=0$; hence, (2.14) and (2.15) are satisfied. Moreover,

$$D\phi_\varepsilon(x)=\frac{1}{\varepsilon}D\psi\left(\frac{x}{\varepsilon}\right)h_s(x)+\psi\left(\frac{x}{\varepsilon}\right)Dh_s(x);$$

therefore,

$$\|D\phi_\varepsilon(x)\|\leqq\begin{cases}\frac{L}{\varepsilon}(c\|x\|+\|h_s(0)\|)+c, & \frac{\varepsilon}{2}\leqq\|x\|<\varepsilon,\\ c \quad \text{otherwise}\end{cases}$$

and since $\|h_s(0)\|=\|\phi(0)-h_s(0)\|<\varepsilon$ the above inequality implies (2.16). Finally, for any sequences $\{\varepsilon_i\}\subset R^+$ and $\{x_i\}\subset Q$ with $\varepsilon_i\to 0$ and $x_i\to x$ it follows by the uniform convergence of $h_{s(\varepsilon_i)}$ to ϕ that $h_{s(\varepsilon_i)}(x_i)\to\phi(x)$. Since ψ is uniformly bounded on R^n, there exist subsequences $\{\varepsilon_i'\}\subset\{\varepsilon_i\}$ and $\{x_i'\}\subset\{x_i\}$ and a constant $0\leqq\sigma\leqq 1$ such that $\psi(x_i'/\varepsilon_i')\to\sigma$. Therefore $\phi_{\varepsilon_i'}(x_i')\to\sigma\phi(x)$ and so (2.17) is fulfilled. □

We are now in a position to prove Theorem 2.4. We shall establish only statement (ii). The proof of the first part of the theorem follows using similar arguments and is left to the reader.

Proof of Theorem 2.4 (ii). Suppose that the system (2.1) is stabilizable by means of the Lipschitzian feedback law $v=\phi(x_1)$ and $\phi(0)=0$. Then there is a smooth Lyapunov function V defined on a compact neighborhood U of R^{n_1} and a positive and strictly increasing continuous function $c: R^+\to R^+$ such that $c(0)=0$ and

$$f(V)(x_1,\phi(x_1))\leqq -c(\|x_1\|)$$

for any $x_1\in U$ [24]. Let $K>0$ and $Q\subset U$ as defined in Lemma 2.5, where ϕ is the above Lipschitzian stabilizer, which is defined on U and it takes values on R^{n_2}. Consider positive constants L_1 and L_2 such that $\|\phi(x_1)\| < L_1/2$ for $x_1 \in Q$ and $\|Df_1(x)\| < L_2$ for every x belonging to

$$N \overset{\text{def}}{=} \{x \in R^n \colon x_1 \in Q, \|x_2\| < L_1\}.$$

For any $r > 0$, such that the sphere $S(0, r)$ of radius r around $0 \in R^n$ is contained to N, consider a positive $l = l(r) < 1$ such that the region

$$N_l \overset{\text{def}}{=} \{x \in R^n \colon \|x_1\| \leqq l, \|x_2\| \leqq 2lK\}$$

is contained to $S(0, r)$. Let $\varepsilon = \varepsilon(r)$ with $\varepsilon < l$ and

$$\varepsilon < \frac{L_2^{-1}}{2} \min \{\|DV(x_1)\|^{-1} c(\|x_1\|) \colon x_1 \in Q, \|x_1\| \geqq l\}.$$

According to Lemma 2.5 there exists a smooth map ϕ_ε such that $\phi_\varepsilon(0) = 0$,

$$\|\phi(x_1) - \phi_\varepsilon(x_1)\| < \varepsilon \quad \forall x_1 \in Q, \quad \|x_1\| \geqq \varepsilon, \tag{2.18}$$

$$\|D\phi_\varepsilon(x_1)\| < K \quad \forall x_1 \in Q. \tag{2.19}$$

Then for any $x_1 \in Q$ with $\|x_1\| > l$ we get

$$\begin{aligned} f_1(V)(x_1, \phi_\varepsilon(x_1)) &\leqq |f_1(V)(x_1, \phi_\varepsilon(x_1)) - f_1(V)(x_1, \phi(x_1))| + f_1(V)(x_1, \phi(x_1)) \\ &\leqq L_2 \|DV(x_1)\| \, \|\phi(x_1) - \phi_\varepsilon(x_1)\| - c(\|x_1\|) \qquad (2.20) \\ &\leqq -\tfrac{1}{2} c(\|x_1\|) < 0. \end{aligned}$$

We define

$$\Phi_r(x) = V(x_1) + \tfrac{1}{2} \|x_2 - \phi_\varepsilon(x_1)\|^2 \quad (\varepsilon = \varepsilon(r)). \tag{2.21}$$

Obviously, Φ_r is continuously differentiable and positive definite for $x \in N$. Similar to Theorem 4 of [37], we consider smooth functions $h_1, \cdots, h_{n_2}$ of $x_1 \in R^{n_1}$, $\hat{x}_2, x_2 \in R^{n_2}$ such that

$$f_1(V)(x_1, \hat{x}_2) - f_1(V)(x_1, x_2) = \sum_{i=1}^{n_2} (\hat{x}_2 - x_2)_i h_i(x_1, \hat{x}_2, x_2)$$

and let $m_r \colon N \to R^{n_2}$ be a real map with components

$$(m_r(x))_i = \begin{cases} D(\phi_\varepsilon)_i f_1(x_1, x_2) - h_i(x_1, \phi_\varepsilon(x_1), x_2) - (x_2 - \phi_\varepsilon(x_1))_i & \text{for } x \notin N_l, \|x_1\| > l, \\ D(\phi_\varepsilon)_i f_1(x_1, x_2) - h_i(x_1, \phi_\varepsilon(x_1), x_2) - (x_2 - \phi_\varepsilon(x_1))_i & \\ \quad - \dfrac{f_1(V)(x_1, \phi_\varepsilon(x_1))}{\|x_2 - \phi_\varepsilon(x_1)\|^2} (x_2 - \phi_\varepsilon(x_1))_i & \text{for } x \notin N_l, \|x_1\| \leqq l, \\ 0 \quad \text{for } x \in N_l. \end{cases}$$

Note that for $x \in N - N_l$ and $\|x_1\| < l$ it holds that $\|x_2\| > 2Kl$ and by (2.19) that $\|\phi_\varepsilon(x_1)\| < Kl$; therefore,

$$\|x_2 - \phi_\varepsilon(x_1)\| \geqq \|x_2\| - K\|x_1\| > Kl \neq 0.$$

For each $x \in N - N_l$ we find

$$(f(\Phi_r) + g(\Phi_r) m_r)(x) = \begin{cases} f_1(V)(x_1, \phi_\varepsilon(x_1)) - \|x_2 - \phi_\varepsilon(x_1)\|^2, & \|x_1\| > l, \\ -\|x_2 - \phi_\varepsilon(x_1)\|^2, & \|x_1\| \leqq l. \end{cases}$$

Consequently, by (2.20) it follows that

$$(f(\Phi_r)+g(\Phi_r)m_r)(x)<0 \quad \forall x\in N-N_l.$$

Since f_1, h_i, and ϕ_ε are continuously differentiable, vanishing at zero, and because of (2.18) and (2.19), there is a positive constant M which is independent of r, satisfying

$$\|m_r(x)\|<M\|x\|+c_r, \quad \forall x\in N,$$

where

$$c_r=\sup\left\{\frac{|f_1(V)(x_1,\phi_\varepsilon(x_1))|}{\|x_2-\phi_\varepsilon(x_1)\|}, x\in N-N_l, \|x_1\|<l\right\} \qquad (\varepsilon=\varepsilon(r)).$$

Note that

$$\frac{|f_1(V)(x_1,\phi_\varepsilon(x_1))|}{\|x_2-\phi_\varepsilon(x_1)\|}\leqq\frac{L_2(1+K)\|DV(x_1)\|}{\|x_2\|-K\|x_1\|}\|x_1\|<L_2\left(1+\frac{1}{K}\right)\|DV(x_1)\|$$

for $x\in N-N_l$ and $\|x_1\|<l$. Since $DV(x_1)$ is continuously vanishing at zero and $\lim l(r)=\lim\varepsilon(r)=0$ as $r\to 0$ it follows that $c_r\to 0$ as $r\to 0$. Finally, for every sufficiently small r, condition (1.7a) holds. Indeed, otherwise there would exist sequences $\{x_{r_i}\}\subset\partial N$ and $\{y_{r_i}\}\subset S[0,r_i]$ such that $x_{r_i}\to x=(x_1',x_2')'\in\partial N$ and $\Phi_{r_i}(x_{r_i})\leqq\Phi_{r_i}(y_{r_i})$. Since $\varepsilon(r)\to 0$ it follows from (2.17) that there exist subsequences $\{x_{r_i}'\}\subset\{x_{r_i}\}$ and $\{y_{r_i}'\}\subset\{y_{r_i}\}$ and real constants σ_1 and σ_2 satisfying

$$V(x_1)+\tfrac{1}{2}\|x_2-\sigma_1\phi(x_1)\|^2=\lim\Phi_{r_i}(x_{r_i}')\leqq\lim\Phi_{r_i}(y_{r_i}')=V(0)+\tfrac{1}{2}\sigma_2^2\|\phi(0)\|^2=0,$$

and so $x_1=x_2=0$, a contradiction. Similarly, we can show (1.7b). We conclude that $\{\Phi_r\}$ as defined in (2.21) is a control Lyapunov family, which satisfies the small control property and so the proof of Theorem 2.4(ii) is completed. □

The following theorem is a consequence of the theory that we have developed and summarizes some useful properties concerning the stability behavior of (2.13).

THEOREM 2.6. *Consider the system* (2.13) *and suppose that* (2.1) *is stabilizable by means of the feedback law* $v=\phi(x_1)$, $\phi(0)=0$.

(i) *If* ϕ *is continuous, then* (2.13) *is practically stabilizable at zero.*

(ii) *Furthermore, if* ϕ *is Lipschitz continuous, then* (2.13) *is practically stabilizable at zero, where the corresponding family of the smooth stabilizers* $\{k_r\}$ *satisfies* (1.8).

(iii) *If, in addition,* ϕ *is continuously differentiable, then* (2.13) *is asymptotically stabilizable at zero, by means of a feedback law, which is smooth for* $x\neq 0$ *near zero and continuous at zero.*

3. Lyapunov functions and uncontrollable eigenvalues. In this section we examine the relationship between various types of Lyapunov conditions and the position on the complex field of the uncontrollable eigenvalues l_u of the linearization

$$(A,B)=(Df(0),g(0)) \tag{3.1}$$

of (1.1) at zero.

It is well known that a necessary condition for stabilization by means of a differentiable feedback is that the real parts of the uncontrollable eigenvalues of (3.1) are nonpositive:

$$\operatorname{Re} l_u\leqq 0. \tag{3.2}$$

If the differentiability assumptions for the stabilizing feedback are relaxed, the previous statement is no longer true [37], [38]. It turns out that, in general, the existence of a control Lyapunov function does not imply (3.2).

Next we establish that (3.2) is a necessary condition for the existence of a control Lyapunov function Φ satisfying the *small control property* and further there are positive constants k_1 and k_2 such that for any x near zero there holds

$$k_1\|x\|^2 \leqq \Phi(x) \leqq k_2\|x\|^2. \tag{3.3}$$

Similarly, condition (3.2) is necessary for the existence of a control Lyapunov family $\{\Phi_r\}$ satisfying the small control property and, in addition, there are constants k_1 and k_2 such that for any x near zero we have

$$k_1\|x\|^2 \leqq \Phi_r(x) \leqq k_2\|x\|^2, \qquad r>0. \tag{3.4}$$

Note that (3.3) is fulfilled, if for instance we assume that Φ is smooth and the matrix $D^2\Phi(0)$ is positive definite:

$$D^2\Phi(0)>0. \tag{3.5}$$

Indeed, since $0\in R^n$ is a local minimum for Φ, we have

$$\Phi(x)=x'D^2\Phi(0)x+O(x^2), \quad x \text{ near zero}$$

$$|O(x^2)|/\|x\|^2\to 0 \quad \text{as } x\to 0,$$

which in conjunction with (3.5) implies (3.3).

Moreover, we note that if there exists a control Lyapunov function, which satisfies (3.3), then (1.1) is strongly asymptotically stabilizable [38], namely, the solution $x(t, x_0)$ of (1.2) satisfies the additional property

$$\|x(t, x_0)\| \leqq m\|x_0\|, \quad m=\sqrt{k_2k_1^{-1}} \quad \forall t \geqq 0. \tag{3.6}$$

Similarly, the existence of a control Lyapunov family which satisfies (3.4) implies strong practical stabilization. That means there is a family of constants $\varepsilon_r>0$ with $\varepsilon_r\to 0$ as $r\to 0$ such that the trajectory $x_r(t, x_0)$ of (1.3) satisfies the additional property

$$\|x_r(t, x_0)\| \leqq m\|x_0\|+\varepsilon_r, \quad \forall t \geqq 0.$$

The following theorem generalizes Proposition 7 of [38].

THEOREM 3.1. *Assume that the system* (1.1) *is strongly asymptotic stabilizable by means of the feedback law* $u=k(x)$, *and there exists a strictly increasing continuous function* $c: R^+\to R^+$, $c(0)=0$ *such that*

$$\|(B-g(x))k(x)\| \leqq \|x\|c(\|x\|) \tag{3.7}$$

for x near zero. Then all the uncontrollable eigenvalues l_u *of* (A, B) *have nonpositive real parts. The same conclusion follows if we assume that* (1.1) *is strongly practical stabilizable by means of the family of feedback laws* $\{k_r\}$ *such that for almost all r and for x near zero there holds*

$$\|(B-g(x))k_r(x)\| \leqq \|x\|(c(\|x\|)+c_r),$$

where $c: R^+\to R^+$, $c(0)=0$ *is a strictly increasing continuous function and* $\{c_r\}$ *is a family of positive constants converging to zero as* $r\to 0$.

Outline of the proof. Suppose that (1.1) is strongly asymptotic stabilized by the feedback k, and let us, on the contrary, assume that the claim is not true. Then there is a linear change of coordinates such that the closed-loop system becomes

$$\begin{pmatrix}\dot{x}_1\\ \dot{x}_2\end{pmatrix}=\begin{pmatrix}A_1 & A_2\\ 0 & A_3\end{pmatrix}\begin{pmatrix}x_1\\ x_2\end{pmatrix}+\begin{pmatrix}B_1k(x)\\ 0\end{pmatrix}+\begin{pmatrix}F_1(x, k(x))\\ F_2(x, k(x))\end{pmatrix}, \tag{3.8}$$

where A_1, A_2, A_3, and B_1 are constant matrices, all the eigenvalues of A_3 have strictly positive real parts, and therefore there exist constants L and $p>0$ such that

$$\|e^{-A_3 t}\| \leqq L\, e^{-pt} \quad \forall t \geqq 0. \tag{3.9}$$

Furthermore, (3.7) is invariant under any linear change of coordinates; hence, there are positive constants M_1 and M_2 satisfying

$$\|F_2(x, k(x))\| < M_1(\|x\|^2 + \|x\| c(M_2\|x\|)). \tag{3.10}$$

The second equation of (3.8) is $\dot{x}_2 = A_3 x_2 + \hat{F}_2(x)$, $\hat{F}_2(x) = F_2(x, k(x))$ and its solution is written as

$$\begin{aligned} x_2(t, x_{20}) &= e^{A_3 t} x_{20} + \int_0^t e^{A_3(t-s)} \hat{F}_2(x(s, x_{20}))\, ds \\ &= e^{A_3 t}\left(x_{20} + \left(\int_0^{+\infty} - \int_t^{+\infty} \right) e^{-A_3 s} \hat{F}_2(x(s, x_{20}))\, ds \right). \end{aligned}$$

By (3.6) and (3.10) it follows that

$$\|\hat{F}_2(x(t, x_0))\| \leqq M_1(m^2\|x_0\|^2 + m\|x_0\| c(M_2 m\|x_0\|)) \tag{3.11}$$

and therefore it can be easily established that the integral

$$I(x_0) = \int_0^{+\infty} e^{-A_3 s} \hat{F}_2(x(s, x_0))\, ds$$

exists. In particular, (3.9), (3.11), and stability imply that there are positive C_1 and C_2 such that

$$\|I(x_0)\| \leqq C_1\|x_0\|^2 + C_2\|x_0\| \int_0^{+\infty} e^{-ps} c(M_2 m\|x_0\|)\, ds < +\infty.$$

Let

$$\theta(x_{20}) = x_{20} + \int_0^{+\infty} e^{-A_3 s} \hat{F}(x(s, x_0))\, ds|_{x_{10}=0}.$$

Then, similar to Proposition 7 of [38], condition (3.11) implies that $\theta(x_{20}) \neq 0$ for $x_{20} \neq 0$ near zero, but on the other hand by (3.9) and the fact that

$$0 \leftarrow x_2(t, x_{20}) = e^{A_3 t}\theta(x_{20}) - \int_t^{\infty} e^{-A_3 s} \hat{F}_2(x(sm\, x_0))\, ds \quad \text{as } t \to +\infty,$$

we have $\theta(x_{20}) = 0$, a contradiction. The rest of the theorem follows by applying a similar argument as before and as those given in Proposition 7 of [38]. □

The previous result leads to the following theorem.

THEOREM 3.2. *Assume that* (1.1) *satisfies the* L.C. (P.L.C.) *and the corresponding Lyapunov function* $\Phi(\Phi_r, r>0)$ *satisfies condition* (3.3) ((3.4), *respectively*). *Then all the uncontrollable eigenvalues* l_u *of* (A, B) *have nonpositive real parts provided that, in addition, one of the three following assumptions hold*:

(A_1) *The Lyapunov function* $\Phi(\Phi_r, r>0)$ *satisfies the small control property.*

(A_2) *The Lyapunov function* $\Phi(\Phi_r, r>0)$ *satisfies the bounded control property and there is a real constant* $c>0$ *such that* $\|B - g(x)\| < c\|x\|^2$, *for* x *near zero.*

(A_3) *The control term* g *is constant.*

Proof. Suppose that Φ is a control Lyapunov function and, in addition, one of the assumptions (A_1), (A_2), and (A_3) holds. Then by Theorem 1.2 there exists a stabilizer $k(x)$ which is smooth for $x \neq 0$ near zero and further $k(x) \to k(0) = 0$, if (A_1) holds, and $k(x)$ is bounded locally around zero, provided that (A_2) is satisfied. Then it is straightforward to see that there exists a nonnegative continuous real function $c_1 : R^n \to R$, $c_1(0) = 0$, such that $\|(B - g(x))k(x)\| < \|x\| c_1(x)$ for x near zero. Therefore (3.7) is fulfilled with $c : R^+ \to R^+$, $c(0) = 0$ being any strictly increasing continuous function such that $c_1(x) \leqq c(\|x\|)$ for x near zero. Hence by Theorem 3.1 all the uncontrollable eigenvalues of (A, B) have nonpositive real parts. □

Another important situation arises if we assume that there exists a control Lyapunov function satisfying the following strong type of L.C.

Assumption 3.3. Suppose that there exist a neighborhood N of $0 \in R^n$, a positive-definite smooth function $\Phi : N \to R$ and positive constants r_1 and r_2 such that for each $x \neq 0$ near zero there is a vector $u \in R^l$ satisfying the inequalities

$$\|u\| < r_1 \|x\|,$$
$$f(\Phi)(x) + g(\Phi)(x)u < -r_2 \|x\|^2.$$

THEOREM 3.4. *The following statements are equivalent*:

(i) *The system* (1.1) *satisfies Assumption* 3.3.

(ii) *All the uncontrollable eigenvalues of* (A, B) *have strictly negative real parts*: $\operatorname{Re} l_u < 0$.

(iii) *The system* (1.1) *is* (*locally*) *exponentially stabilizable by means of a linear feedback.*

Proof. (i)$\Rightarrow$(ii) Assumption 3.3 and smoothness of Φ, f, and g imply that for any $x \neq 0$ near zero there is a vector u such that $\|u\| < r_1 \|x\|$, which satisfies

$$\begin{aligned} D\Phi(x)(A + Bu) &= f(\Phi)(x) + g(\Phi)(x)u + D\Phi(x)(A - f(x)) + D\Phi(x)(B - g(x))u \\ &\leqq -r_2 \|x\|^2 + O(x^2) < 0; \end{aligned}$$

therefore, the pair (A, B) satisfies the L.C. Invoking Theorem 5 of [38] it follows that $\operatorname{Re} l_u < 0$. The implications (ii)$\Rightarrow$(iii) and (iii)$\Rightarrow$(i) are well-known consequences of the last inequality (see, for instance, [38], [18]). □

4. Numerical examples.

Example 4.1. Consider the system

$$\text{(4.1)} \qquad \begin{pmatrix} \dot w_1 \\ \dot w_2 \\ \dot w_3 \end{pmatrix} = \begin{pmatrix} w_1^3 - w_1^2(w_2 + w_3) \\ w_2 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ (w_1 - w_2)^3 \\ 0 \end{pmatrix} u_1 + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} u_2.$$

The above system has the form (1.9), where $x_1 = w_1$, $x_2 = (w_2, w_3)'$, $x = (x_1, x_2')'$, $f_1(x) = w_1^3 - w_1^2(w_2 + w_3)$, $f_2(x) = (w_2, 0)'$, $u = (u_1, u_2)'$, and

$$g_2(x) = \begin{pmatrix} (w_1 - w_2)^3 & 0 \\ 0 & 1 \end{pmatrix}.$$

Let $r = (1, -w_3 + w_1)'$. Then the set

$$\text{(4.2)} \qquad M = \{x \in R^3, x_2 = \phi(x_1)\} \quad \text{where } \phi(x_1) = (x_1, x_1)'$$

is asymptotically stable with respect to (2.12). This follows easily by evaluating the time derivative $\dot W$ of the Lyapunov function

$$W(x) = \tfrac{1}{2} \|x_2 - \phi(x_1)\|^2$$

along the trajectories of (2.12). Indeed, we find

$$\dot{W}(x) = -(w_1 - w_2)^4 - (w_1 - w_3)^2 < 0 \quad \forall x \notin M.$$

Moreover, $W(x) = 0$ for $x \in M$ and $W(x) > 0$ otherwise; therefore, M is asymptotically stable with respect to (2.12). Finally the law $v = \phi(x_1)$, where ϕ is defined in (4.2), asymptotically stabilizes (2.1). Indeed, let $V(x_1) = \frac{1}{2}x_1^2$. Then we find

$$\dot{V}(x_1) = DV(x_1)f(x_1, \phi(x_1)) = -x_1^4 < 0 \quad \forall x_1 \neq 0.$$

Therefore (4.1) satisfies the assumptions of Theorem 2.2 and according to Theorem 1.2 it is asymptotically stabilizable at zero. Note that its linearization at the origin is uncontrollable and contains a positive eigenvalue; hence, it cannot be stabilized by a smooth law. Moreover, it is straightforward to see that according to Theorem 3.1 the system (4.1) cannot be strongly stabilized.

Example 4.2. Consider the planar system

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} x_1 + x_2^3 \\ 0 \end{pmatrix} + u \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \tag{4.3}$$

which has the form (2.13) with $f_1(x_1, x_2) = x_1 + x_2^3$. Note that the system $\dot{x}_1 = f_1(x_1, v) = x_1 + v^3$ is stabilized by the continuous law $v = -(2x_1)^{1/3}$. Therefore, by statement (i) of Theorem 2.6 the system (4.3) is practically stabilizable at zero by means of a family of smooth feedback laws. The system cannot be locally stabilized by a smooth law since its linearization contains a positive eigenvalue. It cannot even be strongly stabilized, since its control term is constant (Theorem 3.1).

Example 4.3. Consider the system $\ddot{x} + \phi_1(x, \dot{x}) + u(\dot{x} + \phi_2(x)) = 0$, $x, u \in R$, or equivalently,

$$\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = \begin{pmatrix} x_2 \\ \phi_1(x_1, x_2) \end{pmatrix} + u \begin{pmatrix} 0 \\ x_2 + \phi_2(x_1) \end{pmatrix}, \tag{4.4}$$

where we have assumed that $x_1 \phi_2(x_1) > 0$, for any $x_1 \neq 0$ near zero. Obviously, the system (4.4) has the form (1.9), and the function $\Phi(x_1, x_2) = x_1^2 + (x_2 + \phi_2(x_1))^2$ is a control Lyapunov function for (4.4), whereas the law $v = -\phi_2(x_1)$ asymptotically stabilizes (2.1) at zero. Therefore, by Theorems 1.2 and 2.1(i) the system (4.4) is asymptotically stabilizable by means of a feedback law which is smooth in a neighborhood of $0 \in R^2$, except possibly at zero. Note also that (2.3) is fulfilled with $a = 2$, $r = -1$, and $W = (x_2 + \phi_2(x_1))^2$; therefore, the function Φ satisfies the bounded control property provided that $D\phi_2(0) \neq 0$. In that case zero is exponentially stable with respect to (2.4), and according to Theorem 2.1(ii) the system (4.1) is asymptotically stabilized by a bounded feedback law.

Example 4.4. Consider the system

$$\begin{pmatrix} \dot{w}_1 \\ \dot{w}_2 \\ \dot{w}_3 \end{pmatrix} = \begin{pmatrix} -w_1^3 \\ w_2^3 + w_2 w_1(w_1 + w_2) + w_3(w_1^2 + w_2^2) \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} u. \tag{4.5}$$

The above system has the form (2.13), where $x_1 = (w_1, w_2)'$, $x_2 = w_3$, $x = (x_1', x_2)'$, and $f_1(x) = (-w_1^3, w_2^3 + w_1 w_2(w_1 + w_2) + w_3(w_1^2 + w_2^2))'$. Let

$$\phi(x_1) = \begin{cases} -2w_2 - \dfrac{w_1 w_2^2}{(w_1^2 + w_2^2)} & \text{for } x_1 \neq 0, \\ 0 & \text{for } x_1 = 0. \end{cases}$$

The map ϕ is differentiable for $x_1 \neq 0$ and, in addition, its derivative is uniformly bounded on R^2; hence, ϕ satisfies a Lipschitz condition on R^2. Furthermore, if we apply the feedback law $v = \phi$ in (2.1), the resulting system is

$$\begin{pmatrix} \dot{w}_1 \\ \dot{w}_2 \end{pmatrix} = \begin{pmatrix} -w_1^3 \\ -w_2^3 - w_2 w_1^2 \end{pmatrix}, \tag{4.6}$$

and obviously $0 \in R^2$ is asymptotically stable with respect to the above system. This follows easily if we compute the time derivative of the Lyapunov function $V(x) = w_1^2 + w_2^2$ along the trajectories of (4.6). Therefore, according to Theorem 2.6(ii) the system (4.7) is practically stabilizable at $0 \in R^3$ by means of a family of smooth feedback laws satisfying (1.8).

REFERENCES

[1] D. Aeyels, *Stabilization of a class of nonlinear systems by a smooth feedback control*, Systems Control Lett., 5 (1985), pp. 289–294.

[2] A. Andreini, A. Bacciotti, and G. Stephani, *Global stabilizability of homogeneous vector fields of odd degree*, Systems Control Lett., 10 (1989), pp. 251–256.

[3] Z. Artstein, *Stabilization with relaxed controls*, Nonlinear Anal. Methods Theory Appl., 7 (1983), pp. 1163–1173.

[4] A. Bacciotti and P. Boieri, *Linear stabilizability of planar nonlinear systems*, Math. Control Signals Systems, 3 (1990), pp. 183–193.

[5] S. P. Banks, *Stabilizability of finite- and infinite-dimensional bilinear systems*, IMA J. Math. Control Inform., 3 (1986), pp. 255–271.

[6] B. R. Barmish, M. J. Corless, and G. Leitmann, *A new class of stabilizing controllers for uncertain dynamical systems*, SIAM J. Control Optim., 21 (1983), pp. 246–255.

[7] W. M. Boothby and R. Marino, *Feedback stabilization of planar nonlinear systems*, Systems Control Lett., 12 (1989), pp. 87–92.

[8] R. W. Brockett, *Asymptotic stability and feedback stabilization*, in Differential Geometric Control Theory, R. W. Brockett, R. S. Millman, and H. J. Sussmann, eds., Birkhauser, Boston, 1983, pp. 181–191.

[9] C. I. Byrnes and A. Isidori, *New results and counterexamples in nonlinear feedback stabilization*, Systems Control Lett., 12 (1989), pp. 437–442.

[10] ———, *Local stabilization of minimum-phase nonlinear systems*, Systems Control Lett., 11 (1988), pp. 9–17.

[11] M. J. Corless and G. Leitmann, *Continuous state feedback guaranteeing uniform boundedness for uncertain dynamical systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 1139–1144.

[12] P. E. Crouch, *Spacecraft attitude control and stabilization: applications of geometric control theory*, IEEE Trans. Automat. Control, 29 (1984), pp. 321–333.

[13] W. P. Dayawansa and C. F. Martin, *Asymptotic stabilization of two dimensional real-analytic systems*, Systems Control Lett., 12 (1989), pp. 205–211.

[14] P. O. Gutman, *Stabilizing controllers for bilinear systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 917–922.

[15] W. Hahn, *Stability of Motion*, Springer-Verlag, Berlin, 1967.

[16] H. Hermes, *On the synthesis of a stabilizing feedback control via Lie algebraic methods*, SIAM J. Control Optim., 18 (1980), pp. 352–361.

[17] V. Jurdjevic and J. P. Quinn, *Controllability and stability*, J. Differential Equations, 28 (1978), pp. 381–389.

[18] N. Kalouptsidis and J. Tsinias, *Stability improvement of nonlinear systems by feedback*, IEEE Trans. Automat. Control, 29 (1984), pp. 364–367.

[19] M. Kawski, *Stabilization of nonlinear systems in the plane*, Systems Control Lett., 12 (1989), pp. 169–176.

[20] J. Kurzweil, *On the inversion of Lyapunov's second theorem on stability of motion*, American Mathematical Society Translations, Series 2, Vol. 24, American Mathematical Society, Providence, RI, pp. 19–77, 1956.

[21] J. P. La Salle and S. Lefschetz, *Stability by Liapunov's Direct Method with Applications*, Academic Press, New York, 1961.

[22] K. K. Lee and A. Arapostathis, *Remarks on smooth feedback stabilization of nonlinear systems*, Systems Control Lett., 10 (1988), pp. 41-44.

[23] R. Marino, *Feedback stabilization of single-input nonlinear systems*, Systems Control Lett., 10 (1988), pp. 201-206.

[24] J. L. Massera, *Contributions to stability theory*, Ann. of Math. (2), 64 (1956), pp. 182-206. Erratum in Ann. of Math. (2), 68 (1958), p. 202.

[25] S. M. Nikolsky, *A Course of Mathematical Analysis*, Mir, Moscow, 1977.

[26] N. Rouche, P. Habets, and M. Laloy, *Stability Theory by Liapunov's Direct Method*, Springer-Verlag, Berlin, New York, 1977.

[27] M. Slemrod, *Stabilization of bilinear control systems with applications to nonconservative problems in elasticity*, SIAM J. Control Optim., 16 (1978), pp. 131-141.

[28] E. D. Sontag, *Conditions for abstract nonlinear regulation*, Inform. and Control, 51 (1981), pp. 105-127.

[29] ———, *A Lyapunov-like characterization of asymptotic controllability*, SIAM J. Control Optim., 21 (1983), pp. 462-471.

[30] ———, *Smooth stabilization implies coprime factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 435-443.

[31] ———, *A "universal" construction of Artstein's theorem on nonlinear stabilization*, Systems Control Lett., 13 (1989), pp. 117-123.

[32] ———, *Feedback stabilization of nonlinear systems*, in Proc. MTNS '89, 1989.

[33] E. D. Sontag and H. J. Sussmann, *Remarks on continuous feedback*, in Proc. IEEE Conference on Decision and Control, Albuquerque, December 1980.

[34] ———, *Further comments on the stabilizability of the angular velocity of a rigid body*, Systems Control Lett., 12 (1989), pp. 213-217.

[35] H. J. Sussmann, *Subanalytic sets and feedback control*, J. Differential Equations, 31 (1979), pp. 31-52.

[36] H. J. Sussmann and P. V. Kokotovic, *The peaking phenomenon and the global stabilization of nonlinear systems*, Report 89-07, SYCON-Rutgers Center for Systems and Control, Rutgers University, New Brunswick, NJ, March 1989.

[37] J. Tsinias, *Sufficient Lyapunovlike conditions for stabilization*, Math. Control Signals Systems 2 (1989), pp. 343-357.

[38] J. Tsinias, *Stabilization of Affine in Control Nonlinear Systems*, Nonlinear Anal. Methods Theory Appl., 12 (1988), pp. 1283-1296.

[39] J. Tsinias and N. Kalouptsidis, *Output feedback stabilization*, IEEE Trans. Automat. Control, (1990) to appear.

[40] A. J. Van der Schaft, *Stabilization of Hamiltonian Systems*, J. Nonlinear Anal. Methods Theory Appl., 10 (1986), pp. 1021-1035.

[41] M. Vidyasagar, *Decomposition techniques for large-scale systems with nonadditive interactions: stability and stabilizability*, IEEE Trans. Automat. Control, 25 (1980), pp. 773-779.

[42] ———, *On the stabilization of nonlinear systems using state detection*, IEEE Trans. Automat. Control, 25 (1980), pp. 504-509.

[43] F. W. Wilson, *Smoothing derivatives of functions and applications*, Tech. Report 66-3, Brown University, Providence, RI, 1966.

[44] ———, *The structure of the level Surfaces of a Lyapunov function*, J. Differential Equations, 3 (1967), pp. 323-329.

[45] T. Yoshizawa, *Stability theory by Liapunov's Second Method*, Mathematical Society of Japan, Tokyo, 1966.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 474-492, March 1991

SECOND-ORDER HAMILTON-JACOBI EQUATIONS IN INFINITE DIMENSIONS*

PIERMARCO CANNARSA†§ AND GIUSEPPE DA PRATO‡§

Abstract. Some second-order Hamilton-Jacobi equations connected to stochastic optimal control problems for infinite-dimensional systems driven by a white noise are studied. A direct method to prove existence and uniqueness of mild solutions is developed. Then this solution is identified as the value function of the related stochastic control problem, and a feedback formula for optimal controls is derived.

Key words. Hamilton-Jacobi equations, stochastic optimal control, dynamic programming, viscosity solutions, white noise, infinite dimensions

AMS(MOS) subject classifications. 49C10, 49A60, 93E20

1. Introduction. Second-order Hamilton-Jacobi equations in infinite dimensions have been studied by several authors in connection with the stochastic optimal control of distributed parameter systems; see [Lecture Notes in Mathematics, Vol. 1390, Springer-Verlag, Berlin, 1989] and the references quoted therein. Most of the works on this subject concern systems governed by stochastic partial differential equations driven by a Hilbert space-valued Wiener process. In this paper, we focus our attention on the case of stochastic systems that are driven by a white noise. For such problems fewer results are available in the literature.

In order to explain the context we have in mind, let X be a separable Hilbert space, and consider the problem of minimizing:

$$J_\varepsilon(t,x;z)=\mathscr{E}\left\{\int_t^T\left[g(y_\varepsilon(s))+\frac{1}{2}|z(s)|^2\right]ds+\phi(y(T))\right\} \tag{1.1}$$

over all controls $z\in M_W^2(t,T;X)$ satisfying $|z(s)|\leqq R$ almost surely for all $s\in[t,T]$. Here ε, R, and T are given positive numbers and $g,\phi:X\to\mathbf{R}$ are bounded uniformly continuous functions. In (1.1), y_ε is the mild solution of the stochastic differential equation

$$\begin{aligned} &dy_\varepsilon(s)=(Ay_\varepsilon(s)+F(y_\varepsilon(s))+z(s))\,dt+\sqrt{\varepsilon}\,dW(s), \qquad t\leqq s\leqq T,\\ &y_\varepsilon(t)=x, \end{aligned} \tag{1.2}$$

where W is a cylindrical Wiener process (or *white noise*) on a probability space $(\Omega,\mathscr{F},P)$. Moreover, $M_W^2(t,T;X)$ denotes the space of the X-valued processes $x(s)$ that are adapted to W and satisfy

$$\mathscr{E}\left(\int_t^T|x(s)|^2\,ds\right)<\infty.$$

As is well known, the dynamic programming approach to problem (1.1), (1.2) consists of studying the *value function* V_ε, defined as

$$V_\varepsilon(t,x)=\inf\{J_\varepsilon(t,x;z): z\in M_W^2(t,T;X), |z(s)|\leqq R \text{ a.s. } \forall s\in[t,T]\}. \tag{1.3}$$

* Received by the editors December 4, 1989; accepted for publication (in revised form) April 25, 1990.

† Dipartimento di Matematica, Università di Pisa, Via F. Buonarroti, 2, 56127 Pisa, Italy. This research was partially supported by the Italian National Project M.P.I. Equazioni Differenziali e Calcolo delle Variazioni.

‡ Scuola Normale Superiore, Piazza dei Cavalieri, 7, 56126 Pisa, Italy.

§ This research was partially supported by the Italian National Project M.P.I. Equazioni di Evoluzione e Applicazioni Fisico-Matematiche.

The function $u_\varepsilon(t, x) = V_\varepsilon(T-t, x)$ is related to the Hamilton-Jacobi-Bellman equation

$$\begin{aligned}&\frac{\partial u}{\partial t}=\frac{\varepsilon}{2}\operatorname{Tr}(u_{xx})+\langle Ax+F(x), u_x\rangle-H(u_x)+g(x) \quad \text{in }]0, T[\times X,\\ &u(0,x)=\phi(x),\end{aligned} \tag{1.4}$$

where

$$H(p)=\begin{cases}\frac{1}{2}|p|^2 & \text{if } |p|\leqq R,\\ R|p|-\frac{R^2}{2} & \text{if } |p|\geqq R.\end{cases} \tag{1.5}$$

The main goal of this paper is to develop a direct method of solution for equation (1.4). By "direct method" we mean a method that makes no use of the control theoretic interpretation of problem (1.4). More precisely, we will solve the above problem as an initial value problem for a semilinear parabolic equation. Then, after having identified the direct solution of (1.4) as the value function (1.3), we can transfer information from a partial differential equation context into a variational setting, by deriving a feedback formula for optimal controls.

We now explain the main ideas of our method. As in finite dimensions, we first consider the linear problem

$$\begin{aligned}&\frac{\partial u}{\partial t}=\frac{\varepsilon}{2}\operatorname{Tr}(u_{xx})+\langle Ax, u_x\rangle \quad \text{in }]0, T[\times X,\\ &u(0,x)=\phi(x)\end{aligned} \tag{1.6}$$

whose solution can be represented by the probabilistic formula

$$u(t,x)=\mathscr{E}\left[\phi\left(e^{tA}x+\sqrt{\varepsilon}\int_0^t e^{(t-s)A}\,dW(s)\right)\right]=:(T_t\phi)(x). \tag{1.7}$$

Indeed, when A is self-adjoint, strictly negative and A^{-1} is nuclear, it is shown in [7] that (1.7) is the unique *classical* solution of problem (1.6). This result is recalled and improved in § 3 of this paper, by proving the uniform convergence of some finite-dimensional approximations of (1.7).

Then, we define a *mild* solution of (1.4) as a solution of the integral equation

$$u(t,\cdot)=T_t\phi+\int_0^t T_{t-s}(\langle F, u_x(s,\cdot)\rangle-H(u_x(s,\cdot))+g)\,ds. \tag{1.8}$$

We solve (1.8) by fixed-point arguments in a space of functions which are C^1 in x for $t>0$ and satisfy a suitable blowup condition in zero, (see § 4).

In order to apply the above result to problem (1.1), (1.2) we have to identify the function $u(T-t, x)$ as the value function $V_\varepsilon(t, x)$. This could be done by standard verification techniques, computing the Itô differential $du(T-t, y_\varepsilon(t))$, if u were sufficiently smooth and the covariance of W had finite trace. To overcome this difficulty, we study a suitable finite-dimensional approximation of (1.4), for which the smoothness of the solution u_n is well known. We then apply the Itô formula to u_n and pass to the limit as $n\to\infty$. To make this procedure rigorous, it is essential to show that u_n converges to u, uniformly on the bounded sets of $[\tau, T]\times X$ for all $0<\tau<T$ (see Theorem 4.5).

The techniques of this paper could be easily arranged to study the equation

$$(1.9)\qquad \begin{aligned} &\frac{\partial u}{\partial t}=\frac{\varepsilon}{2}\operatorname{Tr}(Qu_{xx})+\langle Ax+F(x),u_x\rangle-H(u_x)+g(x)\quad \text{in }]0,T[\times X,\\ &u(0,x)=\phi(x),\end{aligned}$$

where Q is a self-adjoint positive nuclear operator in X. This problem is related to the optimal control of a system driven by a "genuine" Wiener process. Unlike (1.4), for which no other result seems available in the literature, equation (1.9) has been considered by several authors. In [1], problem (1.9) is studied with $F=0$ and assuming g and ϕ to be convex. In [9], the case of $A=0$ is treated by the theory of abstract Wiener spaces. Note that, even though the equation considered in [9] looks like (1.4), it is indeed equivalent to (1.9). The general equation (1.9) is solved in [5] by using a probabilistic formula like (1.7). In this case, however, we do not get C^1 regularity, but only differentiability in some special directions related to Q.

We conclude this Introduction by recalling that several results on viscosity solutions are available today for Hamilton–Jacobi equations in infinite dimensions (see [4] for first-order equations). Second-order equations have been treated by Lions in [10]–[12]. In this theory, the existence of solutions to (1.9), for weakly continuous data, is usually obtained by variational methods based on the representation formula (1.3).

2. Preliminaries. Let X be a separable Hilbert space, with norm $|\cdot|$. For any $R>0$, we set

$$B_R=\{x\in X;\ |x|\leqq R\}.$$

For any $x,y\in X$ we denote by $x\otimes y$ the operator defined by

$$x\otimes y\cdot z=\langle y,z\rangle x.$$

Let Y be another Hilbert space. We denote by $C_b(X,Y)$ the Banach space of all bounded uniformly continuous mappings $\phi:X\to Y$ endowed with the sup norm $\|\cdot\|_0$. Likewise, $C_b^h(X,Y)$, $h=0,1,2,\cdots$, endowed with the natural norm $\|\cdot\|_h$, is the set of all the mappings $\phi:X\to Y$ which are h times Fréchet differentiable and such that the kth derivative $\phi^{(k)}$ is uniformly continuous and bounded for all $k\leqq h$. Moreover, we set $C_b^h(X,\mathbf{R})=C_b^h(X)$.

For any $\phi\in C_b(X)$ we denote by ω_ϕ a continuity modulus of ϕ, i.e., continuous function $\omega_\phi:[0,\infty[\to[0,\infty[$ satisfying $\omega_\phi(0)=0$ and such that $|\phi(x)-\phi(y)|\leqq\omega_\phi(|x-y|)$ for all $x,y\in X$. It is well known that any function $\phi\in C_b(X)$ possesses a concave continuity modulus.[1]

Lip (X,Y) is the space of all Lipschitz continuous and bounded functions from X to Y, endowed with the norm

$$\|\phi\|_1=\sup\left\{\frac{|\phi(x)-\phi(y)|}{|x-y|};\ x,y\in X;\ x\neq y\right\}+\|\phi\|_0.$$

Throughout the whole paper we fix a complete orthonormal system in X, denoted by $\{e_k\}_{k\in N}$. We define the projection Π_n of X onto the span of $\{e_1,e_2,\cdots,e_n\}$ as follows:

$$(2.1)\qquad \Pi_n=\sum_{k=1}^{n}e_k\otimes e_k\quad\forall n\in\mathbf{N}.$$

[1] Indeed, set $\omega(t)=\sup\{|\phi(x)-\phi(y)|;\ |x-y|\leqq t\}$. Then ω is a nondecreasing subadditive continuity modulus for ϕ. So we can check that the concave envelope of ω has the required properties.

Now, let $\{\alpha_k\}$ be a sequence of positive real numbers. Then there exists a unique self-adjoint operator A in X such that $Ae_k = -\alpha_k e_k$. As is well known, A is densely defined and closed with domain

$$D(A) = \left\{x \in X : \sum_{k=1}^{\infty} \alpha_k^2 \langle x, e_k \rangle^2 < \infty\right\}.$$

Moreover, since A is negative, A generates an analytic semigroup e^{tA} in X and

$$e^{tA}x = \sum_{k=1}^{\infty} e^{-t\alpha_k} \langle x, e_k \rangle \tag{2.2}$$

for all $x \in X$.

Consider now a complete probability space $\{\Omega, \mathcal{F}, \mathbf{P}\}$ and a sequence $\{\beta_k\}$ of standard one-dimensional Brownian motions, mutually independent. We denote by $W^n(t)$ the n-dimensional Brownian motion given by

$$W^n(t) = \sum_{k=1}^{n} \beta_k(t) e_k \tag{2.3}$$

for all $t \geqq 0$. We set

$$W_A^n(t) = \sum_{k=1}^{n} e_k \int_0^t e^{-\alpha_k(t-s)}\, d\beta_k(s), \tag{2.4}$$

$$W_A(t) = \sum_{k=1}^{\infty} e_k \int_0^t e^{-\alpha_k(t-s)}\, d\beta_k(s). \tag{2.5}$$

We note that $W_A^n(t)$ is the stochastic convolution

$$W_A^n(t) = \int_0^t e^{(t-s)A}\, dW^n(s).$$

In general, (2.5) is meaningless since the series in the right-hand side may not converge. The following proposition shows that it becomes meaningful, under some restrictions on the sequence $\{\alpha_k\}$.

PROPOSITION 2.1. *Assume that*

$$\sum_{k=1}^{\infty} \frac{1}{\alpha_k} < \infty. \tag{2.6}$$

Then, the series in (2.5) *converges in* $L^2(\Omega, \mathcal{F}, \mathbf{P}; H)$ *for all* $t \geqq 0$. *Moreover,* $W_A(t)$ *is a Gaussian process with mean zero and covariance operator* Q_t *given by*

$$Q_t x = \sum_{k=1}^{\infty} \frac{1 - e^{-2\alpha_k t}}{2\alpha_k} \langle x, e_k \rangle e_k, \qquad x \in X. \tag{2.7}$$

Proof. For all $t \geqq 0$, we have

$$\sum_{k=1}^{\infty} \mathscr{E}\left[\int_0^t e^{-\alpha_k(t-s)}\, d\beta_k(s)\right]^2 = \sum_{k=1}^{\infty} \int_0^t e^{-2\alpha_k s}\, ds = \sum_{k=1}^{\infty} \frac{1 - e^{-2\alpha_k t}}{2\alpha_k},$$

which is finite in view of (2.6). Therefore, the series in (2.5) converges in X for all $t > 0$ almost surely to a Gaussian process $W_A(t)$. In order to prove (2.7) it suffices to remark that, for all $x, y \in X$, we have

$$\mathscr{E}\langle W_A(t), x \rangle \langle W_A(t), y \rangle = \sum_{k=1}^{\infty} \langle x, e_k \rangle \langle y, e_k \rangle \int_0^t e^{-2\alpha_k(t-s)}\, ds. \qquad \square$$

Remark 2.2. If (2.6) is fulfilled, we write the stochastic convolution (2.5) as follows:

$$W_A(t)=\int_0^t e^{(t-s)A}\,dW(s),$$

where $W(t)=\sum_{k=1}^{\infty} e_k\beta_k(t)$ is usually interpreted as a *white noise.*

In order to show that $W_A(t)$ has continuous trajectories, we will strengthen (2.6), assuming

$$\sum_{k=1}^{\infty} \alpha_k^{2\sigma-1}<\infty \tag{2.8}$$

for some $\sigma\in\,]0,\frac{1}{2}[$. We set

$$q(t)=\sum_{k=1}^{\infty}\frac{1-e^{-2\alpha_k t}}{2\alpha_k}.$$

We note that (2.8) yields

$$q(t)\leqq Mt^{2\sigma} \tag{2.9}$$

for all $t\geqq 0$ and some constant $M>0$.

PROPOSITION 2.3. *Assume* (2.8). *Then* $W_A(t)$ *has* α-*Hölder continuous trajectories for all* $\alpha\in\,]0,\sigma[$.

Proof. Since $\{\beta_k\}$ are independent, we have

$$\begin{aligned}\mathscr{E}|W_A(t)-W_A(s)|^2&=\sum_{k=1}^{\infty}\mathscr{E}\left[\int_0^t e^{-\alpha_k(t-\rho)}\,d\beta_k(\rho)\right]^2+\sum_{k=1}^{\infty}\mathscr{E}\left[\int_0^s e^{-\alpha_k(s-\rho)}\,d\beta_k(\rho)\right]^2\\&\quad-2\sum_{k=1}^{\infty}\mathscr{E}\left[\int_0^t e^{-\alpha_k(t-\rho)}\,d\beta_k(\rho)\int_0^s e^{-\alpha_k(s-\rho)}\,d\beta_k(\rho)\right]\\&=\sum_{k=1}^{\infty}\int_0^t e^{-2\alpha_k\rho}\,d\rho+\sum_{k=1}^{\infty}\int_0^s e^{-2\alpha_k\rho}\,d\rho-2\sum_{k=1}^{\infty}\int_0^s e^{-\alpha_k(t+s-2\rho)}\,d\rho\end{aligned}$$

for all $t\geqq s\geqq 0$. Now, by changing the variable ρ with $t+s-2\rho$, we obtain

$$\mathscr{E}|W_A(t)-W_A(s)|^2=q(t)+q(s)+2\left[q\left(\frac{t-s}{2}\right)-q\left(\frac{t+s}{2}\right)\right] \tag{2.10}$$

for all $t\geqq s\geqq 0$.

Next, note that, since $q(t)-q(s)\leqq q(t-s)$, (2.9) yields $q\in C^{2\sigma}([0,\infty[)$. Therefore, there exists $C>0$ such that

$$\left|q(t)+q(s)-2q\left(\frac{t+s}{2}\right)\right|\leqq C|t-s|^{2\sigma}.$$

From (2.10) it follows that

$$\mathscr{E}|W_A(t)-W_A(s)|^2\leqq C(1+2^{1-2\sigma})|t-s|^{2\sigma}.$$

Since $W_A(t)-W_A(s)$ is a Gaussian process, $\mathscr{E}|W_A(t)-W_A(s)|^{2m}\leqq C'|t-s|^{2m\sigma}$ for all $t,\ s\geqq 0$ and a suitable constant C'. The Kolmogorov test yields the conclusion. □

PROPOSITION 2.4. *Assume* (2.8). *Then, for all* $T>0$,

$$\mathscr{E}\left(\sup_{0\leqq t\leqq T}|W_A(t)|\right)<\infty, \tag{2.11}$$

$$\lim_{n\to\infty}\mathscr{E}\left(\sup_{0\leqq t\leqq T}|W_A^n(t)-W_A(t)|\right)=0. \tag{2.12}$$

Proof. We use the factorization method as in [6]. Set

$$Y(s)=\sum_{k=1}^{\infty} e_k \int_0^s (s-r)^{-\sigma}\, e^{-\alpha_k(s-r)}\, d\beta_k(r),$$

$$Y_n(s)=\Pi_n Y(s)$$

for all $s\geqq 0$. Then, by a straightforward computation,

$$W_A(t)=\frac{\sin \pi\sigma}{\sigma}\int_0^t (t-s)^{\sigma-1}\, e^{(t-s)A}\, Y(s)\, ds, \tag{2.13}$$

$$W_A^n(t)=\frac{\sin \pi\sigma}{\sigma}\int_0^t (t-s)^{\sigma-1}\, e^{(t-s)A_n}\, Y_n(s)\, ds, \tag{2.14}$$

where $A_n=A\Pi_n$. Therefore,

$$W_A(t)-W_A^n(t)=B_n(t)+C_n(t), \tag{2.15}$$

where

$$B_n(t)=\frac{\sin \pi\sigma}{\sigma}\int_0^t (t-s)^{\sigma-1}[e^{(t-s)A}-e^{(t-s)A_n}]\, Y(s)\, ds,$$

$$C_n(t)=\frac{\sin \pi\sigma}{\sigma}\int_0^t (t-s)^{\sigma-1}\, e^{(t-s)A_n}[Y(s)-Y_n(s)]\, ds.$$

We will estimate $B_n(t)$ and $C_n(t)$ separately. After some computations we obtain

$$\mathscr{E}|Y(s)|^2=\sum_{k=1}^{\infty}\int_0^s e^{-2\alpha_k(s-r)}(s-r)^{-2\sigma}\, dr \leqq k_0 \sum_{k=1}^{\infty}\frac{1}{\alpha_k^{1-2\sigma}}=:k_1 \tag{2.16}$$

for some constant $k_0>0$. Since $Y(s)$ is a Gaussian process, (2.16) implies that, for all $m\in \mathbf{N}$,

$$\mathscr{E}|Y(s)|^{2m}\leqq k_m \tag{2.17}$$

for some constant $k_m>0$. Now, by Hölder's inequality and (2.13) it follows that

$$\mathscr{E}\left(\sup_{0\leqq t\leqq T}|W_A(t)|^{2m}\right)\leqq\left(\frac{\sin \pi\sigma}{\sigma}\right)^{2m}\left(\int_0^T s^{2m(\sigma-1)/(2m-1)}\, ds\right)^{2m-1}\int_0^T \mathscr{E}|Y(s)|^{2m}\, ds.$$

Moreover, we conclude that

$$\mathscr{E}\left(\sup_{0\leqq t\leqq T}|B_n(t)|^{2m}\right)\leqq\left(\frac{\sin \pi\sigma}{\sigma}\right)^{2m}\left(\int_0^T s^{2m(\sigma-1)/(2m-1)}\|e^{sA}-e^{sA_n}\|^{2m/(2m-1)}\, ds\right)^{2m-1}$$
$$\cdot\int_0^T \mathscr{E}|Y(s)|^{2m}\, ds.$$

Now, since the semigroup e^{sA} is analytic, $\|e^{sA}-e^{sA_n}\|\to 0$ for all $s>0$ as $n\to\infty$. Thus, in view of (2.17), the dominated convergence theorem yields

$$\lim_{n\to\infty}\mathscr{E}\left(\sup_{0\leqq t\leqq T}|B_n(t)|^{2m}\right)=0 \tag{2.18}$$

provided that $(1-\sigma)2m/(2m-1)<1$. We will now estimate $C_n(t)$: we have

$$\begin{aligned}\mathscr{E}|Y(s)-Y_n(s)|^2&=\sum_{k=n+1}^{\infty}\int_0^s e^{-2\alpha_k(s-r)}(s-r)^{-2\sigma}\, dr\\ &\leqq k_0\sum_{k=n+1}^{\infty}\frac{1}{\alpha_k^{1-2\sigma}}:=k_{1n}\end{aligned} \tag{2.19}$$

for some constant $k_{1n}>0$ such that $\lim_{n\to\infty} k_{1n}=0$. Since $Y(s)-Y_n(s)$ is a Gaussian process, (2.19) implies that, for all $m\in\mathbf{N}$,

$$\mathscr{E}|Y(s)-Y_n(s)|^{2m}\leqq c_m(k_{1n})^m \tag{2.20}$$

for some constant $c_m>0$. Now, by Hölder's inequality we conclude that

$$\mathscr{E}\left(\sup_{0\leqq t\leqq T}|C_n(t)|^{2m}\right)\leqq\left(\frac{\sin\pi\sigma}{\sigma}\right)^{2m}\left(\int_0^T s^{2m(\sigma-1)/(2m-1)}\,ds\right)^{2m-1}$$
$$\cdot\int_0^T \mathscr{E}|Y(s)-Y_n(s)|^{2m}\,ds.$$

Hence

$$\lim_{n\to\infty}\mathscr{E}\left(\sup_{0\leqq t\leqq T}|C_n(t)|^{2m}\right)=0, \tag{2.21}$$

which, along with (2.18), gives the conclusion. □

3. Linear parabolic equations. In this section we study the linear problem:

$$\frac{\partial u}{\partial t}=\frac{\varepsilon}{2}\operatorname{Tr}(u_{xx})+\langle Ax, u_x\rangle \quad \text{in } [0, T]\times X, \qquad u(0,x)=\phi(x), \tag{3.1}$$

where ε is a given positive number and $\phi\in C_b(X)$. Here $A:D(A)\subset X\to X$ is a self-adjoint negative operator in X satisfying, for all $k\in\mathbf{N}$,

$$Ae_k=-\alpha_k e_k, \tag{3.2}$$

with $\alpha_k>0$. In (3.1) the subscript x represents Fréchet partial derivative with respect to x and Tr denotes the trace; i.e.,

$$\operatorname{Tr}(u_{xx})=\sum_{k=1}^{\infty}\langle u_{xx}e_k, e_k\rangle.$$

The following result, proved in [7], states that, for any $\phi\in C_b(X)$, problem (3.1) has a unique classical solution given by

$$u(t,x)=\mathscr{E}(\phi(e^{tA}x+\sqrt{\varepsilon}\;W_A(t)))=:(T_t\phi)(x), \tag{3.3}$$

where $W_A(t)$ is the process defined in (2.5). The method of [7], based on a procedure of Galerkin type, uses the following sequence of semigroups which is proved to converge pointwise to the semigroup defined in (3.3):

$$\mathscr{E}(\phi(e^{tA}\Pi_n x+\sqrt{\varepsilon}\;W_A^n(t)))=:(T_t^n\phi)(x) \tag{3.4}$$

for all $\phi\in C_b(X)$. In this paper we improve the result of [7], by showing that $(T_t^n\phi)(x)$ converges to $(T_t\phi)(x)$ uniformly. We will use this convergence result in § 5. In the following we denote by $\overline{e^{tA}Q_t^{-1}}$ the bounded operator defined by

$$\overline{e^{tA}Q_t^{-1}}\;e_k=\frac{2\alpha_k e^{-t\alpha_k}}{1-e^{-2t\alpha_k}}\,e_k,\qquad k\in\mathbf{N}. \tag{3.5}$$

PROPOSITION 3.1. *Assume* (3.2) *and* (2.6). *Then, for any* $\tau\in\,]0, T[$,

$$\text{(i)}\quad \lim_{n\to\infty}(T_t^n\phi)(x)=(T_t\phi)(x),$$
$$\text{(ii)}\quad \lim_{n\to\infty}(T_t^n\phi)_x(x)=(T_t\phi)_x(x) \tag{3.6}$$

uniformly on the bounded sets of $[\tau, T]\times X$. *Moreover, the function*

$$u:[0,\infty[\times X\to \mathbf{R},\ u(t,x)=(T_t\phi)(x)$$

is continuous. Furthermore, $u(t,\cdot)\in C_b^\infty(X)$ *for all* $t>0$, $u(\cdot,x)\in C^1(]0,\infty[)$ *for all* $x\in D(A)$ *and*

$$u_x(t,x)=\sqrt{\varepsilon}\ \mathscr{E}(\overline{e^{tA}Q_t^{-1}}W_A(t)\phi(e^{tA}x+\sqrt{\varepsilon}\ W_A(t))), \tag{3.7}$$

$$u_{xx}(t,x)=\varepsilon\mathscr{E}(\overline{e^{tA}Q_t^{-1}}\ W_A(t)\otimes\overline{e^{tA}Q_t^{-1}}\ W_A(t)\phi(e^{tA}x+\sqrt{\varepsilon}\ W_A(t))), \tag{3.8}$$

$$|u_x(t,x)|\leqq\sqrt{\varepsilon}\ \rho(t)\|\phi\|_0, \tag{3.9}$$

$$|\mathrm{Tr}\,[u_{xx}(t,x)]|\leqq\varepsilon\rho^2(t)\|\phi\|_0, \tag{3.10}$$

where

$$\rho^2(t)=\sum_{k=1}^{\infty}\frac{2\alpha_k e^{-2t\alpha_k}}{1-e^{-2t\alpha_k}}. \tag{3.11}$$

Finally, (3.1) *is fulfilled for all* $x\in D(A)$ *and* $t>0$.

Proof. First, we note that formulas (3.7) and (3.8) are derived in [7]. We will give a detailed proof of the uniform convergence in (3.6). The proofs of the remaining statements will be only sketched for the reader's convenience.

From (3.3) and (3.4) we obtain

$$\begin{aligned}|(T_t\phi)(x)-(T_t^n\phi)(x)|&\leqq\mathscr{E}|\phi(e^{tA}x+\sqrt{\varepsilon}\ W_A(t))-\phi(e^{tA}\Pi_nx+\sqrt{\varepsilon}\ W_A^n(t))|\\&\leqq\mathscr{E}\omega_\phi(|e^{tA}x-e^{tA}\Pi_nx|+\sqrt{\varepsilon}\ |W_A(t)-W_A^n(t)|)\\&\leqq\omega_\phi(|e^{tA}x-e^{tA}\Pi_nx|+\sqrt{\varepsilon}\ \mathscr{E}|W_A(t)-W_A^n(t)|),\end{aligned} \tag{3.12}$$

where ω_ϕ is a concave continuity modulus for ϕ. Now, since e^{tA} is compact for $t>0$,

$$\lim_{n\to\infty}|e^{tA}x-e^{tA}\Pi_nx|=0 \tag{3.13}$$

uniformly for $(t,x)\in[\tau,T]\times B_R$, $R>0$. Moreover,

$$\mathscr{E}|W_A(t)-W_A^n(t)|\leqq\sqrt{T}\sqrt{\mathscr{E}|W_A(t)-W_A^n(t)|^2}. \tag{3.14}$$

Thus, (3.6)(i) follows from (3.13) and (3.14), in light of Proposition 2.1. Equation (3.6)(ii) follows by a similar argument, using formula (3.7) for $u_x(t,x)$. Next, we have

$$\begin{aligned}|\langle u_x(t,x),e_k\rangle|^2&=\varepsilon\,e^{-2t\alpha_k}\left(\frac{2\alpha_k}{1-e^{-2t\alpha_k}}\right)^2\left|\mathscr{E}\int_0^t e^{-2(t-s)\alpha_k}\,d\beta_k(s)\phi(e^{tA}x+W_A(t))\right|^2\\&\leqq\varepsilon\,e^{-2t\alpha_k}\left(\frac{2\alpha_k}{1-e^{-2t\alpha_k}}\right)^2\mathscr{E}\left|\int_0^t e^{-2(t-s)\alpha_k}\,d\beta_k(s)\right|^2\|\phi\|_0^2\end{aligned} \tag{3.15}$$

and (3.9) follows taking the sum over k. Similarly, (3.8) yields

$$\begin{aligned}|\langle u_{xx}(t,x)e_k,e_k\rangle|&=\varepsilon\,e^{-2t\alpha_k}\left(\frac{2\alpha_k}{1-e^{-2t\alpha_k}}\right)^2\Bigg|\mathscr{E}\\&\quad\cdot\left\{\left[\int_0^t e^{-2(t-s)\alpha_k}\,d\beta_k(s)\right]^2\phi(e^{tA}x+W(t))\right\}\Bigg|\\&\leqq\varepsilon\,e^{-2t\alpha_k}\frac{2\alpha_k}{1-e^{-2t\alpha_k}}\|\phi\|_0,\end{aligned} \tag{3.16}$$

which in turn implies (3.10). □

Remark 3.2. Under the stronger condition (2.8), we have

$$\rho^2(t) \leqq \sup_k \left(\frac{2\alpha_k e^{-t\alpha_k}}{1-e^{-2t\alpha_k}} \right)^2 \cdot q(t) \leqq \frac{L}{t^2} q(t), \tag{3.17}$$

where

$$L = 4 \max_{\alpha \geqq 0} \frac{\alpha e^{-\alpha}}{1-e^{-2\alpha}}. \tag{3.18}$$

Hence, estimate (3.10) and (2.9) yield

$$|\mathrm{Tr}\,[u_{xx}(t,x)]| \leqq \frac{C\varepsilon}{t^{2-2\sigma}}. \tag{3.19}$$

4. Semilinear parabolic equations. Let $T > 0$ and consider the Cauchy problem

$$\begin{aligned} &\frac{\partial u}{\partial t} = \frac{\varepsilon}{2} \mathrm{Tr}\,(u_{xx}) + \langle Ax + F(x), u_x \rangle - H(u_x) + g(x) \quad \text{in } [0, T] \times X, \\ &u(0,x) = \phi(x), \end{aligned} \tag{4.1}$$

where we assume the following, in addition to (3.2) and (2.8),

$$\begin{aligned} &\text{(i)} \quad \phi, g \in C_b(X), \\ &\text{(ii)} \quad H \in \mathrm{Lip}\,(X), H(0) = 0, \\ &\text{(iii)} \quad F \in C_b(X; X). \end{aligned} \tag{4.2}$$

Obviously, the requirement $H(0) = 0$ implies no loss of generality, as we can replace g by $g - H(0)$. We will solve problem (4.1) in the Banach space

$$\Sigma = \{ v \in C_b([0, T] \times X) \colon v_x \in C(]0, T] \times X; X), t^{1-\sigma} v_x \in B(]0, T] \times X; X) \},$$

$$\|v\|_\Sigma = \sup \{ |v(t,x)| + |t^{1-\sigma} v_x(t,x)| \colon (t,x) \in]0, T] \times X \},$$

where σ is defined in (2.8) and $B(]0, T] \times X; X)$ denotes the space of all bounded X-valued functions defined in $]0, T] \times X$.

DEFINITION 4.1. A function $u \in \Sigma$ is called a mild solution of problem (4.1) if u is a solution of the integral equation

$$u(t,\cdot) = T_t\phi + \int_0^t T_{t-s}(\langle F, u_x(s,\cdot)\rangle - H(u_x(s,\cdot)) + g)\, ds \quad \forall t \in [0, T], \tag{4.3}$$

where T_t is the semigroup defined in (3.3).

LEMMA 4.2. *Assume* (2.8) *and let* $\psi :]0, T] \times X \to \mathbf{R}$ *be such that*

(i) $\psi \in C_b([\tau, T] \times X)$ *for all* $\tau \in]0, T]$,

(ii) $|t^{1-\sigma}\psi(t,x)| \leqq K$ *for all* $(t,x) \in]0, T] \times X$ *and some constant* $K > 0$.

Set

$$f(t,\cdot) = \int_0^t T_{t-s}(\psi(s,\cdot))\, ds \quad \forall t \in]0, T]. \tag{4.4}$$

Then $f \in \Sigma$.

Proof. Step 1. $f \in C_b([0, T] \times X)$. Fix $\varepsilon > 0$ and let $\tau \in]0, T]$ such that $K\tau^\sigma \leqq \sigma\varepsilon$. Let $(t,x), (t',x') \in [0, T] \times X$.

We shall consider two cases separately.

Case 1. $t, t' \in [0, \tau]$. Then we have, obviously,

$$|f(t,x)-f(t',x')| \leqq \int_0^t |T_{t-s}(s^{1-\sigma}\psi(s,\cdot))(x)|s^{\sigma-1}\,ds + \int_0^{t'} |T_{t-s}(s^{1-\sigma}\psi(s,\cdot))(x')|s^{\sigma-1}\,ds$$

$$\leqq 2K\int_0^\tau s^{\sigma-1}\,ds \leqq 2\varepsilon.$$

Case 2. $t \in]\tau/2, T[$, $t' \in [\tau, T]$. Then we have

$$|f(t,x)-f(t',x')| \leqq 2\varepsilon + \left|\int_{\tau/2}^t T_{t-s}(\psi(s,\cdot))(x)\,ds - \int_{\tau/2}^{t'} T_{t'-s}(\psi(s,\cdot))(x')\,ds\right|.$$

In view of (ii), standard continuity properties of the integral in the right-hand side imply that there exists $\delta > 0$ such that, if $|t-t'|+|x-x'|<\delta$, then

$$\left|\int_\tau^t T_{t-s}(\psi(s,\cdot))(x)\,ds - \int_\tau^{t'} T_{t'-s}(\psi(s,\cdot))(x')\,ds\right| < \varepsilon.$$

To conclude the reasoning, set $\delta' = \min\{\delta, \tau/2\}$. Then, from the above analysis it follows that $|f(t,x)-f(t',x')| \leqq 3\varepsilon$ provided $|t-t'|+|x-x'|<\delta$. This proves that $f \in C_b([0,T]\times X)$.

Step 2. $t^{1-\sigma}f_x(t,x)$ is bounded on $]0,T]\times X$. First, note that, by (3.9), (2.8), and Remark 3.2, we obtain

$$|(T_{t-s}\psi(s,\cdot))_x| \leqq \frac{C_0K}{(t-s)^{1-\sigma}s^{1-\sigma}}, \qquad 0<s<t, \tag{4.5}$$

where $C_0 = \sqrt{\varepsilon LM}$, L and M being defined in (3.17) and (2.9), respectively. Hence, $f_x(t,x)$ exists for all $t>0$ and $x \in X$. Moreover, by (ii),

$$|t^{1-\sigma}f_x(t,x)| \leqq t^{1-\sigma}C_0K\int_0^t (t-s)^{\sigma-1}s^{\sigma-1}\,ds. \tag{4.6}$$

On the other hand, (4.6) yields the conclusion of Step 2 since

$$\int_0^t (t-s)^{\sigma-1}s^{\sigma-1}\,ds = t^{2\sigma-1}\beta(\sigma,\sigma) \leqq \frac{2^{2(1-\sigma)}}{\sigma}t^{2\sigma-1}, \tag{4.7}$$

where β is the Euler beta function.

Step 3. $f_x \in C_b([t_0,T]\times X; X)$ for all $t_0 \in]0,T[$. Fix $t_0 \in]0,T[$ and $t_0 \leqq t \leqq t' \leqq T$, $x, x' \in X$. Then

$$\begin{aligned} &|f_x(t,x)-f_x(t',x')| \\ &\leqq \left|\int_0^t [(T_{t-s}\psi(s,\cdot))_x(x) - (T_{t'-s}\psi(s,\cdot))_x(x')]\,ds\right| \\ &\quad + \left|\int_t^{t'} (T_{t'-s}\psi(s,\cdot))_x(x)\,ds\right|. \end{aligned} \tag{4.8}$$

Moreover, recalling (4.6), we obtain

$$\left|\int_t^{t'} (T_{t'-s}\psi(s,\cdot))_x(x)\,ds\right| \leqq C_0K\int_t^{t'} (t'-s)^{\sigma-1}s^{\sigma-1}\,ds \leqq C_0\frac{K}{\sigma}t_0^{\sigma-1}(t'-t)^\sigma. \tag{4.9}$$

On the other hand, for all $\eta > 0$ there exists $\tau \in]0, t_0/2]$ such that

$$\begin{aligned}(4.10)\qquad &\left|\int_0^t [(T_{t-s}\psi(s,\cdot))_x(x) - (T_{t'-s}\psi(s,\cdot))_x(x')]\,ds\right| \\ &\leqq 2\eta + \left|\int_t^{t-\tau} [(T_{t-s}\psi(s,\cdot))_x(x) - (T_{t'-s}\psi(s,\cdot))_x(x')]\,ds\right|.\end{aligned}$$

Indeed, it suffices to take τ so that

$$(4.11)\qquad C_0 \frac{K}{\sigma}\left(\frac{t_0}{2}\right)^{\sigma-1} \tau^\sigma < \eta.$$

Finally, the conclusion follows from (4.8)-(4.10) and the uniform continuity of the mapping $(s, t, x) \to (T_{t-s}\psi(s,\cdot))_x(x)$ on $\{(s, t, x): t_0 \leqq t \leqq T,\ \tau \leqq s \leqq t-\tau,\ x \in X\}$. The proof of the lemma is thus complete. □

THEOREM 4.3. *Assume* (3.2), (2.8), *and* (4.2). *Then problem* (4.1) *has a unique mild solution.*

Proof. Suppose first that T is sufficiently small, i.e.,

$$(4.12)\qquad (1+2^{2(1-\sigma)}C_0)\frac{\|F\|_0 + \|H\|_1}{\sigma} T^\sigma \leqq \frac{1}{2},$$

where $C_0 = \sqrt{\varepsilon L M}$, L and M being defined in (3.17) and (2.9), respectively. Define a map Γ on Σ as follows:

$$(4.13)\qquad (\Gamma v)(t,\cdot) = T_t\phi + \int_0^t T_{t-s}(\langle F, v_x(s,\cdot)\rangle - H(v_x(s,\cdot)) + g)\,ds$$

for all $t \in [0, T]$. From Lemma 4.2, it follows that $\Gamma: \Sigma \to \Sigma$. Moreover,

$$(4.14)\qquad |(\Gamma v)(t,x) - (\Gamma z)(t,x)| \leqq \frac{\|F\|_0 + \|H\|_1}{\sigma} T^\sigma \|v-z\|_\Sigma,$$

$$(4.15)\qquad t^{1-\sigma}|(\Gamma v)_x(t,x) - (\Gamma z)_x(t,x)| \leqq C_0(\|F\|_0 + \|H\|_1)\beta(\sigma,\sigma)T^\sigma \|v-z\|_\Sigma,$$

where β is the Euler beta function. Since $\beta(\sigma,\sigma) \leqq 2^{2(1-\sigma)}/\sigma$, (4.13)-(4.15) imply that Γ is a contraction in Σ and the conclusion follows by the contraction mapping principle. Finally, condition (4.12) can be removed by a finite number of iterations of the previous fixed-point argument. □

In the sequel we will consider the following "finite-dimensional" approximation of (4.1):

$$(4.16)\qquad \begin{aligned}&\frac{\partial u_n}{\partial t} = \frac{\varepsilon}{2}\operatorname{Tr}(u_{n,xx}) + \langle A\Pi_n x + \Pi_n F(\Pi_n x), u_{n,x}\rangle - H(u_{n,x}) + g(\Pi_n x),\\ &u_n(0,x) = \phi(\Pi_n x),\end{aligned}$$

which has the integral form

$$(4.17)\quad u_n(t,\cdot) = T_t^n\phi\circ\Pi_n + \int_0^t T_{t-s}^n(\langle F\circ\Pi_n, u_{n,x}(s,\cdot)\rangle - H(u_{n,x}(s,\cdot)) + g\circ\Pi_n)\,ds.$$

THEOREM 4.5. *Assume* (3.2), (2.8), *and* (4.2) *and let* u *and* u_n *be the solutions of* (4.1) *and* (4.16), *respectively. Then, for all* $\tau \in]0, T[$,

$$(4.18)\quad \begin{aligned} &\text{(i)}\quad \lim_{n\to\infty} |u(t,x)-u_n(t,x)|=0,\\ &\text{(ii)}\quad \lim_{n\to\infty} |u_x(t,x)-u_{n,x}(t,x)|=0 \end{aligned}$$

uniformly for $t\in[\tau, T]$ *and* x *in bounded sets of* X.

We first prove the following lemma.

LEMMA 4.6. *Assume* (2.8) *and let* $\psi_n, \psi:]0, T]\times X \to \mathbf{R}$, $n\in N$, *be such that*

(i) $\psi_n, \psi \in C_b([\tau, T]\times X)$ *for all* $\tau\in]0, T]$

(ii) $|t^{1-\sigma}\psi_n(t,x)|\leqq K$, $|t^{1-\sigma}\psi(t,x)|\leqq K$ *for all* $(t,x)\in]0,T]\times X$ *and some constant* $K>0$.

(iii) $\lim_{n\to\infty}\sup\{|t^{1-\sigma}(\psi(t,x)-\psi_n(t,x))|: t\in[0,T], |x|\leqq R\}=0$ *for all* $R>0$.

Set

$$f_n(t,\cdot)=\int_0^t (T_{t-s}\psi_n)(s,\cdot)\,ds, \qquad f(t,\cdot)=\int_0^t (T_{t-s}\psi)(s,\cdot)\,ds.$$

Then, for all $R>0$

$$(4.19)\qquad \lim_{n\to\infty} |f_n(t,x)-f(t,x)|=0,$$

$$(4.20)\qquad \lim_{n\to\infty} t^{1-\sigma}|f_{n,x}(t,x)-f_x(t,x)|=0$$

uniformly for $t\in[0,T]$ *and* $|x|\leqq R$.

Proof. First, we note that $f_n, f\in\Sigma$ in view of Lemma 4.2. Now, fix $R>0$ and let $t\in[0,T]$, $|x|\leqq R$. We have

$$(4.21)\qquad \begin{aligned} |f(t,x)-f_n(t,x)|\leqq &\left|\int_0^t T^n_{t-s}[\psi(s,\cdot)-\psi_n(s,\cdot)]\,ds\right|\\ &+\left|\int_0^t [T_{t-s}-T^n_{t-s}]\psi(s,\cdot)\,ds\right|. \end{aligned}$$

We claim that

$$(4.22)\qquad \lim_{n\to\infty}\left|\int_0^t [T_{t-s}-T^n_{t-s}]\psi(s,\cdot)\,ds\right|=0$$

uniformly for $t\in[0,T]$, $|x|\leqq R$. Indeed, fix $\eta>0$ and let $\tau\in]0,T[$ be such that $(K/\sigma)\tau^\sigma<\eta$. Then, if $0\leqq t\leqq\tau$,

$$(4.23)\qquad \left|\int_0^t [T_{t-s}-T^n_{t-s}]\psi(s,\cdot)\,ds\right|\leqq 2\eta.$$

On the other hand, if $\tau<t\leqq T$, then

$$(4.24)\qquad \begin{aligned} \left|\int_0^t [T_{t-s}-T^n_{t-s}]\psi(s,\cdot)\,ds\right| &\leqq 2\eta+\left|\int_{\tau/2}^{t-\tau/2} [T_{t-s}-T^n_{t-s}]\psi(s,\cdot)\,ds\right|\\ &\quad+\left|\int_{t-\tau/2}^{t} [T_{t-s}-T^n_{t-s}]\psi(s,\cdot)\,ds\right|\\ &\leqq 4\eta+\left|\int_{\tau/2}^{t-\tau/2} [T_{t-s}-T^n_{t-s}]\psi(s,\cdot)\,ds\right| \end{aligned}$$

and (4.22) follows from (3.6). Next, let us show that

$$\lim_{n\to\infty}\left|\int_0^t T^n_{t-s}[\psi(s,\cdot)-\psi_n(s,\cdot)]\,ds\right|=0 \tag{4.25}$$

uniformly for $t\in[0,T]$, $|x|\leqq R$. Recalling (3.4), we obtain

$$\begin{aligned} T^n_{t-s}[\psi(s,\cdot)-\psi_n(s,\cdot)](x)=\mathscr{E}[\psi(s,e^{(t-s)A}\Pi_n x+\sqrt{\varepsilon}\;W^n_A(t-s))\\ -\psi_n(s,e^{(t-s)A}\Pi_n x+\sqrt{\varepsilon}\;W^n_A(t-s))].\end{aligned} \tag{4.26}$$

Moreover, Proposition 2.4 implies that there exists a random variable C, such that

$$|e^{(t-s)A}\Pi_n x+\sqrt{\varepsilon}\;W^n_A(t-s)|\leqq R+C \tag{4.27}$$

for $n\in\mathbf{N}$, $0\leqq s\leqq t\leqq T$ and $|x|\leqq R$. So, in light of hypothesis (iii),

$$\begin{aligned}\lim_{n\to\infty}\sup_{|x|\leqq R,\tau\leqq s\leqq t}|\psi(s,e^{(t-s)A}\Pi_n x+\sqrt{\varepsilon}\;W^n_A(t-s))\\ -\psi_n(s,e^{(t-s)A}\Pi_n x+\sqrt{\varepsilon}\;W^n_A(t-s))|=0\end{aligned} \tag{4.28}$$

almost surely for all $\tau\in]0,T]$. Hence, by the dominated convergence theorem, for all $\tau\in]0,T]$.

$$\lim_{n\to\infty}\left|\int_\tau^t T^n_{t-s}[\psi(s,\cdot)-\psi_n(s,\cdot)](x)\,ds\right|=0 \tag{4.29}$$

uniformly for $|x|\leqq R$ and $\tau\in]0,T]$. Now fix $\eta>0$ and choose τ so that $(K/\sigma)\tau^\sigma<\eta$. Then

$$\left|\int_0^t T^n_{t-s}[\psi(s,\cdot)-\psi_n(s,\cdot)](x)\,ds\right|\leqq 2\eta+\left|\int_\tau^t T^n_{t-s}[\psi(s,\cdot)-\psi_n(s,\cdot)](x)\,ds\right| \tag{4.30}$$

and (4.25) follows from (4.28). Finally, (4.21), (4.22), and (4.25) imply (4.19). Next, we prove (4.20). For all $t\in]0,T]$, $|x|\leqq R$ we have

$$\begin{aligned} t^{1-\sigma}|f_x(t,x)-f_{n,x}(t,x)|\leqq t^{1-\sigma}\left|\int_0^t (T^n_{t-s}[\psi(s,\cdot)-\psi_n(s,\cdot)])_x\,ds\right|\\ +t^{1-\sigma}\left|\int_0^t ([T_{t-s}-T^n_{t-s}]\psi)_x(s,\cdot)\,ds\right|.\end{aligned} \tag{4.31}$$

Now, fix $\eta>0$ and let $\tau\in]0,T[$ be such that $8C_0Kt^\sigma<\sigma\eta$, where $C_0=\sqrt{\varepsilon LM}$, L and M being defined in (3.17) and (2.9), respectively. Then, by (4.5) and (4.7) we conclude that, if $0\leqq t\leqq T$,

$$t^{1-\sigma}\left|\int_0^t ([T_{t-s}-T^n_{t-s}]\psi)_x(s,\cdot)\,ds\right|\leqq 2C_0Kt^{1-\sigma}\int_0^t (t-s)^{\sigma-1}s^{\sigma-1}\,ds\leqq\frac{8}{\sigma}\,C_0Kt^\sigma<\eta.$$

On the other hand, if $\tau<t\leqq T$, we obtain, as in (4.24),

$$\begin{aligned} t^{1-\sigma}\left|\int_0^t ([T_{t-s}-T^n_{t-s}]\psi)_x(s,\cdot)\,ds\right|\\ \leqq 2\eta+t^{1-\sigma}\left|\int_{\tau/2}^{t-\tau/2} ([T_{t-s}-T^n_{t-s}]\psi)_x(s,\cdot)\,ds\right|.\end{aligned} \tag{4.32}$$

So, by (3.6)(ii),

$$\lim_{n\to\infty} t^{1-\sigma}\left|\int_0^t ([T_{t-s}-T^n_{t-s}])_x\psi(s,\cdot)\,ds\right|=0 \tag{4.33}$$

uniformly for $t\in]0, T]$ and $|x|\leqq R$. Next we prove that

$$\lim_{n\to\infty} t^{1-\sigma}\left|\int_0^t (T^n_{t-s}[\psi(s,\cdot)-\psi_n(s,\cdot)])_x\,ds\right|=0 \tag{4.34}$$

uniformly for $t\in]0, T]$ and $|x|\leqq R$. From (3.7) it follows that

$$\begin{aligned}&(T^n_{t-s}[\psi(s,\cdot)-\psi_n(s,\cdot)])_x(x)\\ &\quad=\sqrt{\varepsilon}\ \mathscr{E}\{\overline{e^{(t-s)A}Q^{-1}_{t-s}}\,\Pi_n W_A(t-s)[\psi(s, e^{(t-s)A}\Pi_n x+\sqrt{\varepsilon}\ W^n_A(t-s))\\ &\qquad-\psi_n(s, e^{(t-s)A}\Pi_n x+\sqrt{\varepsilon}\ W^n_A(t-s))]\}.\end{aligned} \tag{4.35}$$

Recalling (4.28) and (2.11), we conclude that, if $t\geqq\tau$, then

$$\begin{aligned}\lim_{n\to\infty}\ \sup_{|x|\leqq R,\tau\leqq s\leqq t}\ &|\overline{e^{(t-s)A}Q^{-1}_{t-s}}\,\Pi_n W_A(t-s)[\psi-\psi_n]\\ &\cdot(s, e^{(t-s)A}\Pi_n x+\sqrt{\varepsilon}\ W^n_A(t-s))|=0\end{aligned} \tag{4.36}$$

almost surely for $|x|\leqq R$, $\tau/2\leqq s\leqq t-\tau/2$. Now, arguing as above, we can prove (4.34) and so (4.20). □

Proof of Theorem 4.5. Set

$$\begin{aligned}G_n(x,p)&=\langle F(\Pi_n x),p\rangle-H(p)+g(\Pi_n x),\\ G(x,p)&=\langle F(\Pi x),p\rangle-H(p)+g(\Pi x)\end{aligned} \tag{4.37}$$

for all x, $p\in X$ and

$$(\Gamma_n v)(t,\cdot)=T^n_t\phi+\int_0^t T^n_{t-s}G(\cdot, v_x(s,\cdot))\,ds\quad \forall v\in\Sigma. \tag{4.38}$$

From Lemma 4.2 it follows that Γ_n maps Σ into Σ. Arguing as in the proof of Theorem 4.3 it follows that

$$\|\Gamma_n v-\Gamma_n z\|_\Sigma\leqq\tfrac{1}{2}\|v-z\|_\Sigma \tag{4.39}$$

provided that T satisfies (4.12). Therefore, Γ_n has a unique fixed-point u_n, which is the unique mild solution of (4.16). Moreover,

$$\|u_n-\Gamma^\mu_n(0)\|_\Sigma\leqq 2^{1-\mu}(T\|g\|_0+\|\phi\|_0), \tag{4.40}$$

where Γ^μ_n denotes the μ-iterate of Γ_n. We claim that for all $\mu\in\mathbf{N}$ and all $R>0$,

$$\lim_{n\to\infty}\Gamma^\mu_n(0)(t,x)=\Gamma^\mu(0)(t,x) \tag{4.41}$$

uniformly for $t\in[\tau, T]$ and $|x|\leqq R$. In fact, (4.41) is true for $\mu=1$, in view of (3.6)(i). Now, suppose that (4.41) holds for $\mu\in\mathbf{N}$. Then the functions

$$\psi(t,x)=G(x,\Gamma^\mu(0))_x(t,x),\qquad \psi_n(t,x)=G(x,\Gamma^\mu_n(0))_x(t,x) \tag{4.42}$$

satisfy the assumptions of Lemma 4.6. Consequently, by (4.19)

$$\lim_{n\to\infty}\Gamma^{\mu+1}_n(0)(t,x)=\Gamma^{\mu+1}(0)(t,x) \tag{4.43}$$

uniformly for $t\in[0, T]$, $|x|\leqq R$. Therefore (4.41) holds for all $\mu\in\mathbf{N}$.

Finally, to prove (4.18)(i) note that for all μ, $n\in\mathbf{N}$

$$\begin{aligned}&|u(t,x)-u_n(t,x)|\\ &\quad\leqq|u(t,x)-\Gamma^\mu(0)(t,x)|+|\Gamma^\mu(0)(t,x)-\Gamma^\mu_n(0)(t,x)|+|\Gamma^\mu_n(0)(t,x)-u_n(t,x)|\\ &\quad\leqq 2^{2-\mu}(T\|g\|_0+\|\phi\|_0)+|\Gamma^\mu(0)(t,x)-\Gamma^\mu_n(0)(t,x)|.\end{aligned} \tag{4.44}$$

Fix $\eta > 0$ and let $\mu_\eta \in \mathbf{N}$ be such that $2^{2-\mu_\eta}(T\|g\|_0 + \|\phi\|_0) < \eta$. Then, (4.44) yields

$$|u(t,x) - u_n(t,x)| \leqq 2\eta + |\Gamma^{\mu_\eta}(0)(t,x) - \Gamma_n^{\mu_\eta}(0)(t,x)|. \tag{4.45}$$

Now, (4.18)(ii) can be easily derived by minor modifications of the above argument (using (3.6)(ii) instead of (3.6)(i) and (4.20) instead of (4.19)). Therefore, the proof is complete. □

5. Application to stochastic optimal control. Let $\{\Omega, \mathcal{F}, \mathbf{P}\}$ be a complete probability space and $\{\beta_k\}$ a sequence of standard one-dimensional Brownian motions, mutually independent. For any $s \geqq 0$ let $\mathcal{F}_t$ be the σ-algebra generated by $\{\beta_k(s)\colon k = 1, 2, \cdots;\ 0 \leqq s \leqq t\}$. Let $M_W^2(t, T; X)$ denote the space of the X-valued processes x such that $x(s)$ in $\mathcal{F}_s$-measurable for all $t \leqq s \leqq T$ and

$$\mathscr{E}\left(\int_t^T |x(s)|^2\, ds\right) < \infty.$$

Consider a stochastic system governed by the state equation

$$y(s) = e^{(s-t)A}x + \int_t^s e^{(s-r)A}[F(y(r)) + z(r)]\, dr + \sqrt{\varepsilon}\ W_A(t,s), \qquad s \geqq t \geqq 0, \tag{5.1}$$

where $x \in X$, A is a self-adjoint operator satisfying (3.2) and (2.8), $F \in \operatorname{Lip}(X, X)$, $z \in M_W^2(t, T; X)$, and $W_A(t,s)$ is defined by

$$W_A(t,s) = \sum_{k=1}^{\infty} e_k \int_t^s e^{-\alpha_k(s-r)}\, d\beta_k(r), \qquad s \geqq t \geqq 0. \tag{5.2}$$

Equation (5.1) can be regarded as the "mild" form of the stochastic differential equation

$$\begin{aligned} &dy(s) = \{Ay(s) + F(y(s)) + z(s)\}\, ds + \sqrt{\varepsilon}\, dW(s), \qquad t \leqq s \leqq T, \\ &y(t) = x, \end{aligned} \tag{5.3}$$

where $W(t)$ is a cylindrical Wiener process (see Remark 2.2).

We now prove the existence of solutions to (5.3) as well as a Galerkin approximation result.

PROPOSITION 5.1. *Assume* (3.2), (2.8) *and let* $F \in \operatorname{Lip}(X, X)$. *Then, for all* $z \in M_W^2(t, T; X)$, *equation* (5.1) *has a unique solution* $y_\varepsilon(\cdot\,; t, x, z)$, *which is continuous with probability one.*

Proof. Let $\Lambda = \{v \in M_W^2(t, T; X)\colon \mathscr{E}(\sup_{t \leqq s \leqq T} |v(s)|^2) < +\infty\}$ and define a map λ on Λ as follows:

$$\lambda(v)(s) = e^{(s-t)A}x + \int_t^s e^{(s-r)A}[F(v(r)) + z(r)]\, dr + \sqrt{\varepsilon}\ W_A(t,s), \qquad t \leqq s \leqq T. \tag{5.4}$$

From Proposition 2.4 it follows that $W_A(t, \cdot) \in \Lambda$. Hence, $\lambda : \Lambda \to \Lambda$. Moreover, λ is a contraction provided that $T - t < 1/\|F\|_1$ and the conclusion follows by standard fixed-point arguments. □

PROPOSITION 5.2. *Assume* (3.2), (2.8) *and let* $F \in \operatorname{Lip}(X, X)$. *Let* $y_{\varepsilon,n}(\cdot\,; t, x, z)$ *be the solution of*

$$\begin{aligned} &dy_{\varepsilon,n}(s) = \{A\Pi_n y_{\varepsilon,n}(s) + \Pi_n F(\Pi_n y_{\varepsilon,n}(s)) + \Pi_n z(s)\}\, ds + \sqrt{\varepsilon}\, dW^n(s), \\ &y_{\varepsilon,n}(t) = \Pi_n x, \qquad\qquad t \leqq s \leqq T, \end{aligned} \tag{5.5}$$

where $z \in M^2_W(t, T; X)$ *and* $W^n(t)$ *is defined in* (2.3). *Then,*

$$\lim_{n\to\infty} \mathscr{E}\left(\sup_{t\leqq s\leqq T} |y_\varepsilon(s; t, x, z) - y_{\varepsilon,n}(s; t, x, z)|\right) = 0 \tag{5.6}$$

for all $x \in X$.

Proof. Let Λ be the space defined in the previous proof and define a map $\lambda_n n$ on Λ as follows:

$$\lambda_n(v)(s) = e^{(s-t)A}\Pi_n x + \int_t^s e^{(s-r)A}\Pi_n[F(\Pi_n v(r)) + \Pi_n z(r)]\, dr + \sqrt{\varepsilon}\,\Pi_n W_A(t, s),$$

$$t \leqq s \leqq T.$$

Then, λ_n is a contraction in Λ uniformly with respect to $n \in \mathbf{N}$, provided that $T - t < 1/\|F\|_1$. Moreover, Proposition 2.4 implies that

$$\lim_{n\to\infty} \mathscr{E}\left(\sup_{t\leqq s\leqq T} |\lambda_n(v)(s) - \lambda(v)(s)|\right) = 0$$

for all $v \in M^2_W(t, T; X)$. The conclusion then follows by the contraction mapping theorem depending on a parameter. □

We will now study the following stochastic optimal control problem.

Given $R > 0$, minimize the cost functional

$$J_\varepsilon(t, x; z) = \mathscr{E}\left\{\int_t^T \left[g(y_\varepsilon(s; t, x, z)) + \frac{1}{2}|z(s)|^2\right] ds + \phi(y_\varepsilon(T; t, x, z))\right\} \tag{5.7}$$

over all controls $z \in M^2_W(t, T; X)$ satisfying $|z(s)| \leqq R$ almost surely for all $s \in [t, T]$.

The *value function* of problem (5.7) is given by

$$V_\varepsilon(t, x) = \inf\{J_\varepsilon(t, x; z) \colon z \in M^2_W(t, T; X), |z(s)| \leqq \mathbf{R}\}. \tag{5.8}$$

The corresponding Hamilton-Jacobi-Bellman equation reads as follows:

$$\begin{aligned} &\frac{\partial v}{\partial t} + \frac{\varepsilon}{2}\operatorname{Tr}(v_{xx}) + \langle Ax + F(x), v_x\rangle - H(v_x) + g(x) = 0 \quad \text{in } [0, T] \times X,\\ &v(T, x) = \phi(x), \end{aligned} \tag{5.9}$$

where H is defined by

$$H(p) = \begin{cases} \frac{1}{2}|p|^2 & \text{if } |p| \leqq R,\\ R|p| - \frac{R^2}{2} & \text{if } |p| \geqq R. \end{cases} \tag{5.10}$$

From Theorem 4.3 we obtain the result below.

THEOREM 5.3. *Assume* (2.1), (2.8), (4.2)(i) *and let* $F \in \operatorname{Lip}(X, X)$. *Then problem* (5.9) *has a unique mild solution, which coincides with the value function* V_ε. *Moreover, for any* $(t, x) \in [0, T] \times X$, *there exists an optimal control for problem* (5.7). *Furthermore, any optimal control* z^* *is related to the corresponding optimal state* y^* *by the feedback formula*

$$z^*(s) = -h\left(\frac{\partial V_\varepsilon}{\partial x}(s, y^*(s))\right), \qquad t \leqq s \leqq T, \tag{5.11}$$

where

$$h(p) = \begin{cases} p & \textit{if } |p| \leqq R,\\ \frac{pR}{|p|} & \textit{if } |p| \geqq R. \end{cases} \tag{5.12}$$

Proof. First, we note that the existence and uniqueness of a mild solution v to (5.9) follows from Theorem 4.3, since the function H defined in (5.10) fulfills (4.2)(ii).

Let us show that $v = V_\varepsilon$. We claim that v satisfies the dynamic programming principle below: for any $t \in]0, T[$, $x \in X$ and $z \in M^2_W(t, T; X)$ such that $|z(s)| \leqq R$ almost surely, we have

$$(5.13)\quad \begin{aligned} &v(t, x) + \frac{1}{2}\mathscr{E}\int_t^T \{|z(s) + v_x(s, y_\varepsilon(s; t, x, z))|^2 - \chi(v_x(s, y_\varepsilon(s; t, z, z)) - R)\}\, ds \\ &\qquad = \mathscr{E}\left\{\int_t^T \left[g(y_\varepsilon(s; t, x, z)) + \frac{1}{2}|z(s)|^2\right] ds + \phi(y_\varepsilon(T; t, x, z))\right\}, \end{aligned}$$

where $\chi(a) = 0$ if $a \leqq 0$ and $\chi(a) = a^2$ if $a \geqq 0$.

Indeed, let u_n be the solution of problem (4.16) with H given by (5.10) and set $v_n(t, x) = u_n(T - t, x)$. We claim that v_n is regular. To show this fact let $\zeta(t, x_1, \cdots, x_n)$ be defined as $\zeta(t, x_1, \cdots, x_n) = v_n(t, x_1 e_1 + \cdots + x_n e_n)$, for all $(t, x_1, \cdots, x_n) \in]0, T[\times \mathbf{R}^n$. Then ζ is a classical solution of the problem

$$\begin{aligned} &\frac{\partial \zeta}{\partial t} + \Delta\zeta - \sum_{i=1}^n [\alpha_i x_i - \langle F(x_1 e_1 + \cdots + x_n e_n), e_i\rangle] \frac{\partial \zeta}{\partial x_i} \\ &\qquad - H\left(\frac{\partial \zeta}{\partial x_1} e_1 + \cdots + \frac{\partial \zeta}{\partial x_n} e_n\right) + g(x_1 e_1 + \cdots + x_n e_n) = 0, \\ &\zeta(T, x_1, \cdots, x_n) = \phi(x_1 e_1 + \cdots + x_n e_n). \end{aligned}$$

So, we can use the Itô formula to differentiate $v_n(s, y_{\varepsilon,n}(s))$ where $y_{\varepsilon,n}(s) = y_{\varepsilon,n}(s; t, x, z)$. Thus, we obtain

$$dv_n(s, y_{\varepsilon,n}(s)) = \frac{\partial v_n}{\partial t}(s, y_{\varepsilon,n}(s))\, ds + \langle dy_{\varepsilon,n}(s), v_{n,x}(s, y_{\varepsilon,n}(s))\rangle + \frac{\varepsilon}{2}\operatorname{Tr}(v_{n,xx}(s, y_{\varepsilon,n}(s)))\, ds.$$

Now, recall (4.16) and (5.5), integrate on $[t, T]$ and take expectation to obtain

$$\begin{aligned} &v_n(t, x) + \frac{1}{2}\mathscr{E}\int_t^T \{|\Pi_n z(s) + v_{n,x}(s, y_{\varepsilon,n}(s))|^2 - \chi(|v_{n,x}(s, y_{\varepsilon,n}(s))| - R\}\, ds \\ &\qquad = \mathscr{E}\left\{\int_t^T \left[g(y_{\varepsilon,n}(s; t, x, z)) + \frac{1}{2}|\Pi_n z(s)|^2\right] ds + \phi(y_{\varepsilon,n}(T; t, x, z))\right\}. \end{aligned}$$

By Proposition 5.2 and Theorem 4.5, we obtain (5.13) in the limit as $n \to \infty$. Next, we note that the following inequality holds:

$$(5.14)\quad |z(s) + v_x(s, y_\varepsilon(s; t, x, z))|^2 - \chi(v_x(s, y_\varepsilon(s; t, x, z)) - R) \geqq 0.$$

Thus, from (5.13) and (5.14) it follows that $v(t, x) \leqq V_\varepsilon(t, x)$.

To prove the reverse inequality, let us consider the closed-loop equation

$$(5.15)\quad \begin{aligned} &y(s) = e^{(s-t)A}x + \int_t^s e^{(s-r)A}[F(y(r)) - h(v_x(r, y(r)))]\, dr + \sqrt{\varepsilon}\, W_A(t, s), \\ &\qquad\qquad T > s \geqq t \geqq 0, \end{aligned}$$

which can be solved by the Schauder fixed-point theorem (see, e.g., [13, Cor. 2.3]). Indeed, from (2.8) it follows that e^{tA} is compact for $t > 0$. Let y^* be a mild solution of (5.15). Taking

$$z(s) = -h(v_x(s, y^*(s))), \tag{5.16}$$

we have the equality in (5.14), and so $v(t, x) \geqq V_\varepsilon(t, x)$ for all $t < T$. Moreover, the choice (5.16) provides an optimal control at (t, x). Finally, the feedback formula (5.11) follows from (5.13) and the fact that $v(t, x) = V_\varepsilon(t, x)$. □

Example 5.4. Let $X = L^2(0, \pi)$ and define

$$\begin{aligned} &D(A) = H^2(0, \pi) \cap H_0^1(0, \pi), \\ &Ax = \frac{\partial^2 x}{\partial \xi^2} \quad \forall x \in D(A), \end{aligned} \tag{5.17}$$

$$F(x)(\xi) = f(x(\xi)), \quad g(x) = \int_0^\pi \alpha(x(\xi))\, d\xi, \quad \phi(x) = \int_0^\pi \beta(z(\xi))\, d\xi,$$

where $f \in C_b^1(\mathbf{R})$, $\alpha, \beta \in C_b(\mathbf{R})$. By inspection, F, g, and ϕ fulfill the hypotheses of Theorem 5.3. As for operator A, (3.2) is satisfied by $\alpha_k = k^2$, and so (2.8) holds true for any $\sigma \in]0, \frac{1}{2}[$. Therefore, the results of this section apply to the following stochastic optimal control problem.

Minimize

$$J_\varepsilon(t, x; z) = \mathscr{E}\left\{\int_t^T \int_0^\pi \left[\alpha(y(s, \xi)) + \frac{1}{2}|z(s, \xi)|^2\right] d\xi\, ds + \int_0^\pi \beta(y(T, \xi))\, d\xi\right\} \tag{5.18}$$

over all controls $z \in M_W^2([t, T]; L^2(0, \pi))$ satisfying $\int_0^\pi |z(s, \xi)|^2\, d\xi \leqq R^2$ almost surely for all $s \in [t, T]$, where the state y is subject to

$$\begin{aligned} &dy(s, \xi) = \left\{\frac{\partial^2}{\partial x^2} y(s, \xi) + f(y(s, \xi)) + z(s, \xi)\right\} ds + \sqrt{\varepsilon}\, dW(s), \\ &y(s, 0) = y(s, \pi) = 0, \qquad s \in [t, T], \\ &y(t, \xi) = x(\xi). \end{aligned} \tag{5.19}$$

Remark 5.5. Let us consider the same problem as in Example 5.4 for an N-dimensional parabolic state equation, i.e., taking $X = L^2([0, \pi]^N)$ and $Ax = \Delta x$, with Dirichlet or Neumann boundary conditions. Then, Theorem 5.3 does not apply. In fact, we can show that $q(t) = t^{1-N/2} o(t)$ (see [7]) and (2.6) is not satisfied. However, if we consider the iterated Laplace operator $A_1 x = (-1)^{m-1}(-\Delta)^m x$, (with Dirichlet boundary conditions), we have $q(t) = t^{1-N/2m} O(t)$ and Theorem 5.3 applies if $N < 2m$.

Remark 5.6. Using Theorem 5.3 and the variational technique of [8] we can characterize the value functions of deterministic optimal control problems as limits, as $\varepsilon \downarrow 0$, of the mild solutions u_ε of (4.1). For example, consider the following problem.

Given $R > 0$, minimize

$$J(t, x; z) = \int_t^T \left[g(y(s; t, x, z)) + \frac{1}{2}|z(s)|^2\right] ds + \phi(y(T; t, x, z)) \tag{5.20}$$

over all controls $z \in L^2(t, T; X)$ satisfying $|z(s)| \leqq R$.

Here $y(\cdot\,; t, x, z)$ is the mild solution of the state equation

$$\begin{aligned} &y'(s) = Ay(s) + F(y(s)) + z(s), \qquad t \leqq s \leqq T, \\ &y(t) = x. \end{aligned}$$

Define the value function of problem (5.20) as

$$V(t, x) = \inf\{J(t, x; z) : z \in L^2(t, T; H), |z(s)| \leqq R\}.$$

Then, for all $(t, x) \in [0, T] \times H$, we can show that

$$|u_\varepsilon(T-t, x) - V(t, x)| \leqq T\omega_g(C\sqrt{\varepsilon}) + \omega_\phi(C\sqrt{\varepsilon}), \tag{5.21}$$

where ω_g (respectively, ω_ϕ) denotes a concave modulus of continuity for g (respectively, ϕ) and

$$C = \sqrt{q(T) - q(t)}\, e^{T\|F\|_1}.$$

REFERENCES

[1] V. Barbu and G. da Prato, *Hamilton-Jacobi Equations in Hilbert Spaces*, Pitman, Boston, 1982.

[2] P. Cannarsa and G. da Prato, *The vanishing viscosity method in infinite dimensions*, Atti Accad. Naz. Lincei, to appear.

[3] ———, *A semigroup approach to Kolmogoroff equations in Hilbert spaces*, Appl. Math. Lett., to appear.

[4] M. G. Crandall and P. L. Lions, *Hamilton-Jacobi equations in infinite dimensions. Part* IV, preprint.

[5] G. da Prato, *Some results on Bellman equation in Hilbert spaces*, SIAM J. Control Optim., 23 (1985), pp. 61-71.

[6] G. da Prato, S. Kwapien, and J. Zabczyk, *Regularity of solutions of linear stochastic equations in Hilbert spaces*, Stochastics, 23 (1987), pp. 1-23.

[7] G. da Prato and J. Zabczyk, *Smoothing properties of transition semigroups in Hilbert spaces*, Stochastics, to appear.

[8] W. H. Fleming, *The Cauchy problem for a nonlinear first order partial differential equation*, J. Differential Equations, 5 (1969), pp. 515-530.

[9] T. Havârneanu, *Existence for the dynamic programming equation of control diffusion processes in Hilbert space*, Nonlinear Anal. Theory Methods Appl., 9 (1985), pp. 619-629.

[10] P. L. Lions, *Viscosity solutions of fully nonlinear second-order equations and optimal stochastic control in infinite dimensions. Part* I: *The case of bounded stochastic evolutions*, Acta Math., 161 (1988), pp. 243-278.

[11] ———, *Viscosity Solutions of Fully Nonlinear Second-Order Equations and Optimal Stochastic Control in Infinite Dimensions. Part* II: *Optimal Control of Zakai's Equation*, Lecture Notes in Mathematics, Vol. 1390, Springer-Verlag, Berlin, 1989.

[12] ———, *Viscosity solutions of fully nonlinear second-order equations and optimal stochastic control in infinite dimensions. Part* III: *uniqueness of viscosity solutions for general second-order equations*, J. Funct. Anal., 86 (1989), pp. 1-18.

[13] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, pp. 493–497, March 1991

CALMNESS AND EXACT PENALIZATION*

J. V. BURKE‡

Abstract. The notion of calmness, which was introduced by Clarke and Rockafellar for constrained optimization, is considered. An equivalence to the technique of exact penalization due to Eremin and Zangwill is established. It is then shown that *calmness* is satisfied on a dense subset of the domain of the optimal value function.

Key words. exact penalization, calmness, constrained optimization

AMS(MOS) subject classifications.49A42, 49D30, 49D37, 90D30

1. Introduction. The notion of calmness was originally formulated by Rockafellar and first appears in the literature of Clarke [3]. Since its appearance it has been recognized as a fundamental concept in optimization theory and consequently many variations on the original definition have been proposed and studied (e.g., see Rockafellar [6]). In general terms, calmness can be described as a basic regularity condition under which we can study the sensitivity properties of certain variational systems. On the other hand, the term "exact penalization" refers to a reduction principle in constrained optimization wherein we replace a constrained optimization problem by an unconstrained optimization problem whose objective is finite-valued on the domain of the original objective function and which under various conditions possesses a common local minimum. The particular reduction technique for exact penalization discussed herein originates in the papers of Eremin [4] and Zangwill [8] (also see Pietrzykowski [5]). We shall establish an equivalence between the notion of calmness and the viability of the Eremin–Zangwill exact penalization procedure for the constrained optimization problem

$$(\mathcal{P}) \qquad \begin{array}{l} \text{minimize } f(x) \\ \text{subject to } g(x) \in C, \end{array}$$

where f is a mapping from the normed linear space X into $\mathbb{R} \cup \{+\infty\}$, g is a mapping from X into the normed linear space Y, and C is a nonempty closed subset of Y. In order to make this statement precise we give the following definitions.

DEFINITION 1.1. Let f, g, X, Y, and C be as in the statement of $\mathcal{P}$ and consider the perturbed problems

$$(\mathcal{P}_u) \qquad \begin{array}{l} \text{minimize } f(x) \\ \text{subject to } g(x) \in C + u, \end{array}$$

where $u \in Y$. Let $\overline{x} \in X$ and $\overline{u} \in Y$ be such that

$$g(\overline{x}) \in C + \overline{u} \text{ and } \overline{x} \in \operatorname{dom}(f) := \{x \in X : f(x) < +\infty\}.$$

*Received by the editors September 8, 1989; accepted for publication (in revised form) March 26, 1990.

‡Department of Mathematics, GN–50, University of Washington, Seattle, Washington 98195. This work was supported in part by National Science Foundation grants DMS–8602399 and DMS–8803206 and by Air Force Office of Scientific Research grant ISSA-860080.

The problem $\mathcal{P}_{\overline{u}}$ is said to be *calm* at $\overline{x}$ if there are constants $\overline{\alpha} \geq 0$ and $\varepsilon > 0$ such that for every pair $(x, u) \in X \times Y$ with $\|x - \overline{x}\| \leqq \varepsilon$ and $g(x) \in C + u$, we have

$$f(x) + \overline{\alpha}\|u - \overline{u}\| \geq f(\overline{x}). \tag{1.1}$$

Here we use the notation $\|z\|$ for the norm of z. The constants $\overline{\alpha}$ and ε are called the *modulus* and *radius* of calmness for $\mathcal{P}_{\overline{u}}$ at $\overline{x}$, respectively.

Remarks. (1) Definition 1.1 varies from the definition given by Clarke [2, Def. 6.4.1] in that the variable u is not required to satisfy $\|u - \overline{u}\| \leq \varepsilon$ in order for inequality (1.1) to hold. In §2, we show that the restriction on the perturbation u is redundant.

(2) Observe that if $\mathcal{P}_{\overline{u}}$ is calm at $\overline{x}$, then $\overline{x}$ is necessarily a local solution to $\mathcal{P}_{\overline{u}}$.

DEFINITION 1.2. Let f be a mapping from the normed linear space X into $\mathbb{R} \cup \{+\infty\}$ and let S be a subset of X. Let $\varepsilon > 0$. We say that $\overline{x} \in S$ is an *ε-local minimum* of f on S if

$$f(\overline{x}) \leq f(x)$$

for all $x \in S$ with $x \in \overline{x} + \varepsilon\mathbb{B}$ where $\mathbb{B}$ is the closed unit ball in X, i.e., $\mathbb{B} := \{v \in X : \|v\| \leqq 1\}$, and $\overline{x} + \varepsilon\mathbb{B} := \{\overline{x} + \varepsilon v : v \in \mathbb{B}\}$.

The main result of this paper can now be stated.

THEOREM 1.1. *Let $\overline{x} \in X$ and $\overline{u} \in Y$ be such that*

$$g(\overline{x}) \in C + \overline{u} \text{ and } \overline{x} \in \operatorname{dom}(f).$$

Then $\mathcal{P}_{\overline{u}}$ is calm at $\overline{x}$ with modulus $\overline{\alpha}$ and radius ε if and only if $\overline{x}$ is an ε-local minimum of

$$P_{\overline{u},\overline{\alpha}}(x) := f(x) + \alpha \operatorname{dom}(g(x)|C + \overline{u}),$$

where we define

$$\operatorname{dist}(z|C + \overline{u}) := \inf\{\|y + \overline{u} - z\| : y \in C\}.$$

Proof. ($\Longrightarrow$) Let $\delta > 0$. Given $x \in \overline{x} + \varepsilon\mathbb{B}$, set $u_x := g(x) - y_x$, where $y_x \in C$ satisfies

$$\|y_x + \overline{u} - g(x)\| \leq \operatorname{dist}(g(x)|C + \overline{u}) + \delta.$$

Note that $g(x) \in C + u_x$ and

$$\|\overline{u} - u_x\| \leq \operatorname{dist}(g(x)|C + \overline{u}) + \delta.$$

Thus, if $\alpha \geq \overline{\alpha}$, we obtain from the calmness hypothesis that

$$\begin{aligned} f(\overline{x}) + \alpha \operatorname{dist}(g(\overline{x})|C + \overline{u}) &= f(\overline{x}) \\ &\leq f(x) + \alpha\|\overline{u} - u_x\| \\ &\leq f(x) + \alpha \operatorname{dist}(g(x)|C + \overline{u}) + \alpha\delta. \end{aligned}$$

Since $\delta > 0$ was chosen arbitrarily the implication is established.

($\Longleftarrow$) Let $u \in Y$ and $x \in \overline{x} + \varepsilon\mathbb{B}$ be such that $g(x) \in C + u$ and $x \in \operatorname{dom}(f)$. Then

$$\begin{aligned} f(\overline{x}) &\leqq f(x) + \overline{\alpha} \operatorname{dist}(g(x)|C + \overline{u}) \\ &= f(x) + \overline{\alpha} \inf\{\|y + \overline{u} - g(x)\| : y \in C\} \\ &\leq f(x) + \overline{\alpha} \inf\{\|y + u - g(x)\| + \|u - \overline{u}\| : y \in C\} \\ &= f(x) + \overline{\alpha} \operatorname{dist}(g(x)|C + u) + \overline{\alpha}\|u - \overline{u}\| \\ &= f(x) + \overline{\alpha}\|u - \overline{u}\|. \end{aligned}$$

Hence $\mathcal{P}_{\overline{u}}$ is calm at $\overline{x}$. □

Remark. The function $P_{u,\alpha}$ defined above is a familiar tool in the mathematical programming literature [2],[4],[5],[8]. For example, in the case where $X := \mathbb{R}^n$, $C := \mathbb{R}^m_- \subset \mathbb{R}^m =: Y$, and Y is endowed with the l_1 norm, we have

$$P_{u,\alpha}(x) = f(x) + \alpha \sum_{i=1}^{m} (g_i(x) - u_i)_+,$$

where g_i and u_i are the ith components of g and u, respectively, and $z_+ := \max\{0, z\}$ for every $z \in \mathbb{R}$.

In the case where Y is finite-dimensional and $g(x)$ is Lipschitz continuous, Clarke [2, Prop. 6.4.3] has shown that the calmness of $\mathcal{P}_{\overline{u}}$ at $\overline{x}$ implies the existence of a constant $\overline{m} > 0$ such that $\overline{x}$ is a local minimum for the function $P_{\overline{u},m}(x)$ for all $m \geqq \overline{m}$. However Clarke's result does not reveal the full extent of the relationship between calmness and exact penalization. In particular, it does not describe the relationship between the parameters $\overline{m}$ and $\overline{\alpha}$ as given in Theorem 1.1.

As previously stated calmness is an important tool in the sensitivity analysis for $\mathcal{P}$. In this regard Theorem 1.1 can be used to study the sensitivity of $\mathcal{P}_u$ to changes in u and to establish multiplier rules for $\mathcal{P}$. These results and others are pursued in Burke [1]. In the remainder of this note we briefly explore two topics directly related to the definitions of calmness and exact penalization as they are employed in Theorem 1.1. In §2 we compare Definition 1.1 to the definition for calmness used by Clarke in [2, Def. 6.4.1]. We conclude in §3 by providing a result that is in the spirit of Clarke's generic calmness result [2, Prop. 6.4.5] indicating the robustness of the notion of calmness.

2. Another formulation of calmness. According to Clarke [2, Def. 6.4.1] in order for $\mathcal{P}_{\overline{u}}$ to be calm at $\overline{x}$ we require that inequality (1.1) be satisfied whenever $\|x - \overline{x}\| \leqq \varepsilon$ and $\|u - \overline{u}\| \leqq \varepsilon$. In the next proposition we show that if g is continuous at $\overline{x}$, then no advantage gained by placing this further restriction on the choice of perturbation u.

PROPOSITION 2.1. *Let f, g, C, X, and Y be as in the statement of problem $\mathcal{P}$. Let $(\overline{x}, \overline{u}) \in X \times Y$ be such that g is continuous at $\overline{x}$ and $g(\overline{x}) \in C + \overline{u}$. If there is an $\overline{\alpha} > 0$ and an $\varepsilon > 0$ such that*

$$f(x) + \overline{\alpha}\|u - \overline{u}\| \geq f(\overline{x})$$

for every pair $(x, u) \in X \times Y$ with $\|u - \overline{u}\| \leq \varepsilon$, $\|x - \overline{x}\| \leq \varepsilon$, and $g(x) \in C + u$, then there is an $\hat{\varepsilon}$ with $0 < \hat{\varepsilon} \leq \varepsilon$ such that $\overline{x}$ is an $\hat{\varepsilon}$-local minimum of $P_{\overline{u},\overline{\alpha}}(x)$, and consequently, $\mathcal{P}_{\overline{u}}$ is calm at $\overline{x}$ with modulus $\overline{\alpha}$ and radius $\hat{\varepsilon}$.

Proof. Let $\delta \in (0, \frac{1}{2})$ and $\alpha \geq \overline{\alpha}$. Since the function $\varphi(x) := \mathrm{dist}(g(x)|C + \overline{u})$ is continuous at $\overline{x}$, there is an $\hat{\varepsilon} \in (0, \varepsilon]$ such that $0 \leq \varphi(x) \leq \frac{\varepsilon}{2}$ whenever $\|x - \overline{x}\| \leq \hat{\varepsilon}$. Now given $x \in \overline{x} + \hat{\varepsilon}\mathbb{B}$, set $u_x := g(x) - y_x$ where $y_x \in C$ satisfies

$$\|y_x + \overline{u} - g(x)\| \leq \mathrm{dist}(g(x)|C) + \delta\varepsilon.$$

Then $g(x) \in C + u_x$ and $\|u_x - \overline{u}\| \leq \varepsilon$. Hence, by hypothesis,

$$\begin{aligned} \mathcal{P}_{\overline{u},\alpha}(\overline{x}) &= f(\overline{x}) \\ &\leq f(x) + \alpha\|u_x - \overline{u}\| \\ &\leq f(x) + \alpha \ \mathrm{dist}(g(x)|C) + \alpha\delta\varepsilon. \end{aligned}$$

Taking the limit as $\delta \downarrow 0$ yields the result. □

Calmness can also be defined independent of the existence of a solution to $\mathcal{P}_u$. This is done by considering the value function for $\mathcal{P}_u$, $V : X \to \mathbb{R} \cup \{\pm\infty\}$, given by

$$V(u) := \inf\{f(x) : g(x) \in C + u\}$$

if $\{x : g(x) \in C + u\} \neq \emptyset$ and $+\infty$ otherwise. $\mathcal{P}_u$ is then said to be calm at $\overline{u}$ if

$$\liminf_{u \to \overline{u}} \frac{V(u) - V(\overline{u})}{\|u - \overline{u}\|} > -\infty. \tag{2.1}$$

In this connection we have the following straightforward extension to Clarke [2, Prop. 6.4.2].

PROPOSITION 2.2. *Let $f, g.C, X$, and Y be as in the statement of $\mathcal{P}$, and let $\overline{u} \in Y$. If* (2.1) *holds at $\overline{u}$, then for any solution $\overline{x}$ to $\mathcal{P}_{\overline{u}}, \mathcal{P}_{\overline{u}}$ is calm at $\overline{x}$.*

3. Calmness is a dense property in finite dimensions. In the case where P_u has the representation

$$\begin{array}{ll} \text{minimize} & f_0(x) \\ \text{subject to} & f_i(x) \le u_i, \ i = 1, \cdots, m \\ & \text{and } x \in D, \end{array}$$

where $f_i : \mathbb{R}^n \to \mathbb{R}$ is locally Lipschitz for each $i = 0, 1, \cdots, m$ and $D \subset \mathbb{R}^n$ is a nonempty closed set, we can employ the results of Clarke [2, §6.4] to show that if $\mathcal{P}_u$ has a finite value for every u near some $\overline{u} \in \mathbb{R}^m$, then for almost every u near $\overline{u}$ (in the sense of Lebesgue measure) $\mathcal{P}_u$ is calm at u. In the spirit of this result we give the following proposition.

PROPOSITION 3.1. *Let f, g, C, X, and Y be as in $\mathcal{P}$. Furthermore, assume that Y is finite-dimensional, f is lower semicontinuous, and g is continuous. If $\overline{u} \in Y$ and $\gamma > 0$ are such that V is bounded on $\overline{u} + \gamma\mathbb{B}$, then $\mathcal{P}_u$ is calm on a dense subset of $\overline{u} + \gamma\mathbb{B}$.*

Proof. With no loss in generality we assume that $\overline{u} = 0$. Let $u \in Y$ be an element of the interior of $\gamma\mathbb{B}$ and $\varepsilon > 0$. We must show that there is a $u_0 \in u + \varepsilon\mathbb{B}$ such that $\mathcal{P}_{u_0}$ is calm. Define $\theta : [0, \gamma - \|u\|] \to \mathbb{R}$ by

$$\theta(\rho) := \inf\{f(x) : g(x) \in C + u + \rho\mathbb{B}\}.$$

The boundedness of $\theta(\rho)$ on $[0, \gamma - \|u\|]$ follows from that of V on $\gamma\mathbb{B}$. Since θ is nonincreasing on $[0, \gamma - \|u\|]$, θ is differentiable at almost every $\rho \in [0, \gamma - \|u\|]$ (in the sense of Lebesgue measure) by Ward [7]. Let ρ_0 be a point of differentiability for θ such that $0 < \rho_0 < \min\{\varepsilon, \gamma - \|u\|\}$. From the definition of θ there is for each $n \in \{1, 2, \cdots\}$ a $u_n \in u + \rho_0\mathbb{B}$ such that

$$V(u_n) - \frac{1}{n} \le \theta(\rho_0). \tag{3.1}$$

Let u_0 be a cluster point of the sequence $\{u_n\}$. Now since f is lower semi continuous and g is continuous, we have that V is also lower semicontinuous, hence, by (3.1), it must be that $V(u_0) = \theta(\rho_0)$. We now show that $\mathcal{P}_u$ is calm at u_0.

Since θ is differentiable at ρ_0, there is a $\delta \in (0, \gamma - \|u\| - \rho_0)$ and an $\alpha > 0$ such that

$$\theta(\rho) - \theta(\rho_0) \ge -\alpha|\rho - \rho_0|$$

whenever $|\rho - \rho_0| < \delta$. Let $\varepsilon_0 \in (0, \min\{\delta, \min\{\varepsilon, \gamma - \|u\|\} - \rho_0\})$, and let $w \in u_0 + \varepsilon_0\mathbb{B}$. Then

$$\begin{aligned} V(w) - V(u_0) &\geqq \theta(\|u - w\|) - \theta(\rho_0) \\ &\geqq -\alpha|\ \|u - w\| - \rho_0|. \end{aligned} \tag{3.2}$$

Now if $\|u - w\| \leq \rho_0$, then $\theta(\|u - w\|) \geq \theta(\rho_0)$ so that

$$V(w) - V(u_0) \geq 0 \geq -\alpha\|w - u_0\|.$$

On the other hand, if $\|u - w\| \geq \rho_0$, then

$$|\ \|u - w\| - \rho_0| \leq \|w - u_0\|,$$

and hence

$$V(w) - V(u_0) \geq -\alpha\|w - u_0\|$$

by (3.2). Therefore

$$V(w) + \alpha\|w - u_0\| \geq V(u_0)$$

for all $w \in u_0 + \varepsilon_0\mathbb{B}$. Consequently, $\mathcal{P}_u$ is calm at u_0. □

The conclusion of Proposition 3.1 is weaker than that of Clarke [2, §6.4] since we do not show that $\mathcal{P}_u$ is calm for almost all u near $\overline{u}$. On the other hand, our result is valid for more general constraints than those considered by Clarke.

Acknowledgment. Just before the final version of this article was sent to the Publisher, we became aware of an important reference on this topic. The reference is Thibault [9]. In this thesis, Thibault establishes certain equivalences between the notions of calmness and exact penalization. These results, although similar to our own, are somewhat different. The results of Thibault are complementary to those presented in this paper.

REFERENCES

[1] J.V. Burke, *An exact penalization viewpoint of constrained optimization*, Report ANL/MCS-TM-95, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, Ill. 1987.

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

[3] ———, *A new approach to Lagrange multipliers*, Math. Oper. Res., 1 (1976), pp.165–174.

[4] I.I. Eremin, *The 'penalty' method in convex programming*, Dokl. Akad, Nauk SSSR, in English. 173 (1966), pp.459–462.

[5] T. Pietrzykowski, *An exact potential method for constrained maxima*, SIAM J. Numer. Anal., 6 (1969), pp.299–304.

[6] R.T. Rockafellar, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Anal., 9 (1985), pp.665–698.

[7] A.J. Ward, *On the differential structure of real functions*, London Math. Soc., 39 (1934), pp.339–362.

[8] W.I. Zangwill, *Nonlinear programming via penalty functions*, Management Sci.13 (1967), pp.344-358.

[9] L. Thibault, *Proprietes des Sous-Differentiels de Fonctions Localement Lipschitziennes Definies sur un Espace de Banach Separable. Applications.*, Ph.D Thesis, Department of Mathematics, University of Montpellier, Montpellier, France, 1975.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 2, p. 498, March 1991

015

ERRATUM:

Controllability of Nonlinear Discrete-Time Systems: A Lie-Algebraic Approach*

BRONISLAW JAKUBCZYK† AND EDUARDO D. SONTAG‡

As stated, Corollary 5.3 is false. The indices are missing in the statement. The correct version is as follows.

COROLLARY 5.3. *If an analytic system with connected* $\mathbb{U}$ *is forward accessible from* x *in* k *steps* (*that is,* $\operatorname{int} A_k^+(x) \neq \emptyset$), *then*

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

Similarly, if it is backward accessible from x *in* k *steps, then*

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

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

$$\bigcup_{k>1} \mathrm{Orb}_{\Delta_k^+}(x) \subset \mathrm{Orb}_L(x).$$

The second statement follows analogously. □

* Received by the editors July 2, 1990; accepted for publication July 5, 1990. SIAM J. Control Optim., 28 (1990), pp. 1-33.

† Institute of Mathematics, Polish Academy of Sciences, Sniadeckich 8, 00-950 Warsaw, Poland.

‡ Rutgers Center for Systems and Control, Department of Mathematics, Rutgers University, New Brunswick, New Jersey 08903. E-mail address: sontag@math.rutgers.edu.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 3, pp. 499-515, May 1991

A NUMERICAL ALGORITHM FOR OPTIMAL FEEDBACK GAINS IN HIGH DIMENSIONAL LINEAR QUADRATIC REGULATOR PROBLEMS*

H. T. BANKS† AND K. ITO†

Abstract. A hybrid method for computing the feedback gains in linear quadratic regulator problems is proposed. The method, which combines use of a Chandrasekhar type system with an iteration of the Newton–Kleinman form with variable acceleration parameter Smith schemes, is formulated to efficiently compute directly the feedback gains rather than solutions of an associated Riccati equation. The hybrid method is particularly appropirate when used with large dimensional systems such as those arising in approximating infinite-dimensional (distributed parameter) control systems (e.g., those governed by delay-differential and partial differential equations). Computational advantages of the proposed algorithm over the standard eigenvector (Potter, Laub–Schur) based techniques are discussed, and numerical evidence of the efficacy of these ideas is presented.

Key words. linear quadratic regulator (LQR) problems, feedback gains, distributed parameter systems, computational algorithm, Chandrasekhar system, Newton–Kleinman scheme, Smith method

AMS(MOS) subject classifications. 65H10, 93B40, 93C20

1. Introduction. A great deal of effort in recent years in control of distributed systems has focused on approximation techniques (for example, see [1]-[6], [8], [11]-[14], [16], [18], [21], [22], [24], [26], [32]) to reduce inherently infinite-dimensional problems to (large) finite-dimensional analogues. Relatively little attention has been given to the development of efficient computational methods for the resulting large but finite-dimensional control problems. In this paper we consider such questions for one classical formulation of the feedback control problem, the well-known linear quadratic regulator (LQR) problem.

There are several approaches we can take in such an endeavor. With the emergence of new computer architectures (vector and parallel), one exciting possibility involves the development of new algorithms to be used with nonsequential computers. While we are currently investigating ideas in this direction , our presentation here reports on some of our efforts to develop better algorithms for use with conventional serial computers.

As is well known, the LQR problem can be reduced to the solution of a matrix Riccati equation in order to construct the feedback gain matrix. The most widely available method for solution of the Riccati equation is the Potter method [30], which is based on obtaining the eigenvectors and eigenvalues of an associated $2n \times 2n$ Hamiltonian system when the underlying dynamical control system is of dimension n. A related, but improved, version involving computation of Schur vectors for the system was proposed by Laub in [27]. While both of these "eigenvector" methods can be used satisfactorily (for a discussion of real advantages offered by the Laub–Schur approach over Potter's method, see [27]) for systems with n relatively small, say

* Received by the editors November 19, 1986; accepted for publication (in revised form) June 6, 1990. This research was supported in part by National Aeronautics and Space Administration grant NAG-1-517, Air Force Office of Scientific Research contracts AFOSR-84-0398 and AFOSR-85-0303, and National Science Foundation grant MCS-8504316. Parts of the research were carried out while the authors were visiting scientists at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, Virginia 23665, which is operated under NASA contracts NAS1-17070 and NAS1-18107.

† Center for Applied Mathematical Sciences, University of Southern California, Los Angeles, California 90089-1113.

$n < 100$, the computational effort (and time) grows like $n^3$and becomes prohibitive for large systems. More recently, the idea of using Chandrasekhar systems ([9], [20], [33, p. 304–310], [36]) when the number of states is large compared to the number of control inputs (exactly the situation in a number of cases where one approximates a distributed system) has been suggested by a number of authors [33, p. 309], [7], [8], [17], [31]. However, as we shall discuss below, there can be numerical difficulties in using the Chandrasekhar approach. On the other hand, it is known that iterative methods such as the Newton algorithm as formulated by Kleinman in [23] can be quite efficiently implemented (even for some large systems) if good initial estimates are provided and if we can efficiently solve the resulting Lyapunov equations. In this presentation, we discuss the formulation and numerical testing of a hybrid method that represents an attempt to combine the good features of the Chandrasekhar approach (growth like n in computational effort) with those of the Newton–Kleinman (quadratic convergence when good initial estimates are provided) along with innovative use of the Smith algorithm for solution of Lyapunov equations.

We expect these ideas to be quite useful in design of control laws for some of the models currently being investigated in connection with large flexible structures as well as in some of the population dispersal and control studies that we are currently pursuing with biologists and ecologists. Some of the large flexible structures involve rather sophisticated distributed parameter models (e.g., see [4], [34]), especially when we wish to include complicated damping mechanisms involving time or spatially related hysteresis [6], [34] or nonlinear effects [19]. For such models, the computational tasks can be rather demanding whether one is carrying out parameter identification [3] or feedback control calculations with traditional eigenvector based methods (the authors of [6] have indicated experiences with runs requiring nine hours of VAX time when using the Potter method for an approximation system with dimension equivalent to $n = 238$).

Our intention is to introduce a problem-oriented algorithm that can be used for large-scale but structured problems within a conventional computational capability rather than to develop a general purpose method for solutions of LQR problems. More precisely, we are interested in computing solutions of the approximating LQR problems for systems governed by partial differential equations. In such a case the dimension n of the states tends to be large while the dimension m of the inputs is fixed under approximation, and moreover the resulting system matrices (A, B) are, in general, sparse. Our hybrid method can be most efficiently carried out for such cases. Of course, this does not imply that our algorithm does not work for general problems, but it simply may not be efficient in some situations.

For our presentation, we assume that one has used his or her favorite approximation scheme (finite-elements, spline, spectral, etc.) to reduce the problems of interest to an LQR problem with a finite-dimensional system. More precisely, throughout our discussions we consider the following LQR problem: minimize the cost functional

$$J(u) = \int_0^\infty \{|Cx(t)|^2 + |u(t)|^2\}\, dt \tag{1.1}$$

subject to the state dynamics

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

Here $A \in R^{n\times n}$, $B \in R^{n\times m}$, and $C \in R^{p\times n}$. (We have, without loss of generality, for our discussions here normalized our problem so that the control term in the cost functional (1.1) appears with a weighting matrix I.) We will assume that (A, B) is stabilizable and (C, A) is detectable [25], [38]. Then the optimal feedback control for the LQR problem involving (1.1) is given by

$$u(t) = -B^T Px(t),$$

where P is the unique nonnegative symmetric solution of the algebraic Riccati equation

$$A^T P + PA - PBB^T P + C^T C = 0. \tag{1.2}$$

In this paper we propose an algorithm which leads to direct calculation of the feedback gain matrix $K = B^T P$ without computation of P. In addition to providing substantial savings in computational time over eigenvector methods, our algorithm requires much less storage and can easily be implemented to take advantage of special structures (e.g., sparsity) in the system matrix A.

To outline the steps in our algorithms, we first recall that the optimal feedback gain K is given by the limit $K = \lim K(t)$ as $t \to -\infty$ of solutions of the Chandrasekhar system [9], [20]

$$\begin{aligned} \dot{K}(t) &= -B^T L^T(t) L(t), \qquad K(0) = 0, \\ \dot{L}(t) &= -L(t)(A - BK(t)), \qquad L(0) = C, \end{aligned} \tag{1.3}$$

where $K \in R^{m \times n}$ and $L \in R^{p \times n}$. In fact (see [20], [37]) $P = \lim \int_t^0 L^T(s) L(s)\, ds$ as $t \to -\infty$. The first step in the proposed hybrid algorithm involves a numerical integration of (1.3) backward in time on an appropriate interval $[-t_f, 0]$. For the second step, the value $K(-t_f)$ obtained through this numerical integration of the Chandrasekhar system is then refined by use of the Newton–Kleinman algorithm [23], if we use $K(-t_f)$ as an initial value K_0 for the Newton method.

To motivate our effort in these two steps, we note that the convergence $K(t) \to K$ as $t \to -\infty$ can be very slow when the eigenvalues of $A - BK$ lie close to the imaginary axis. Moreover, $L = 0$, K_∞ arbitrary are solutions in the asymptotic limit sense to (1.3). That is, if we denote by $f(K, L)$ the right side of system (1.3), then K_∞ arbitrary and $L = 0$ are solutions of $f(K_\infty, L) = 0$. Hence $K(t) \to K_\infty$, $L(t) \to 0$ as $t \to -\infty$ does not, in general, have a unique limit *numerically*. Thus, as is pointed out in [33, p. 316–318], if we are to use the Chandrasekhar approach alone, we need a very accurate numerical solver for (1.3). This can be quite expensive computationally if we are dealing with a large system and/or a stiff system. Hence, we propose to use a second order semi-implicit scheme (described in § 2) for the Chandrasekhar component of our algorithm and take the resulting numerical solution $K(-t_f)$ as a start-up value for the Newton iterations. If this estimate from the Chandrasekhar step is a sufficiently good initial guess, then we can expect to meet the Newton–Kleinman requirements that $A - BK_0$ be a stability matrix and to obtain quadratic convergence in this second component of the algorithm. Thus, the role of the Chandrasekhar step is to provide a good start-up value for the Newton–Kleinman method, not to produce a very accurate solution.

The first step of our hybrid method requires the solution of $n(m+p)$ simultaneous equations, while each iteration of the usual Newton–Kleinman step requires the solution of a Lyapunov equation for the $n \times n$ symmetric estimates of P. However, as we will see below, we can use factorization ideas [20] and the Smith method [35] for Lyapunov equations to reformulate the Newton–Kleinman method as a direct iterative method for the $m \times n$ gain K, thereby providing additional computational advantages. To speed up our calculations and improve convergence in the Smith algorithm, we propose a variable stepsize Smith method to solve the Lyapunov equations as described in § 4 below. In § 2 we outline a numerical scheme for the Chandrasekhar system, while the reformulated Newton-Kleinman iterative procedure to compute directly the gain K is detailed in § 3. Finally, in § 5, we further discuss some advantages and disadvantages of the proposed algorithm and, to illustrate the feasibility of our hybrid approach, report on our experience with two of the several numerical examples on which we have tested the method.

2. Numerical solution of the Chandrasekhar system. We return to consider more closely the Chandrasekhar system

$$\dot{K}(t) = -B^T L^T(t) L(t), \qquad K(0) = 0, \tag{2.1}$$

$$\dot{L}(t) = -L(t)(A - BK(t)), \qquad L(0) = C, \tag{2.2}$$

where $A \in R^{n\times n}$, L, $C \in R^{p\times n}$, and K, $B^T \in R^{m\times n}$. We first observe that the second equation (2.2) is linear in L. In cases where A arises from a discretization or approximation of partial differential equations, equation (2.2) tends to be a stiff system and thus it is advisable to use an implicit numerical scheme. Here we propose the second order Adams-Moulton algorithm [10, p. 235]. A second observation is that the right-hand side of equation (2.1) is independent of K and thus an explicit scheme is appropriate; we propose the second order Adams-Bashforth algorithm [10, p. 226].

These observations leads us to propose the following algorithm for the Chandrasekhar system (2.1), (2.2): Given a step size $h > 0$, approximations K_i and L_i to $K(-ih)$ and $L(-ih)$ are generated by

$$K^{(0)}_{i+1} = K_i + h(\tfrac{3}{2} B^T L_i^T L_i - \tfrac{1}{2} B^T L_{i-1}^T L_{i-1}), \tag{2.3}$$

$$K^{(0)}_{i+1/2} = (K^{(0)}_{i+1} + K_i)/2, \tag{2.4}$$

$$L_{i+1} = L_i + \frac{h}{2}(L_{i+1} + L_i)(A - BK^{(0)}_{i+1/2}), \tag{2.5}$$

$$K_{i+1} = K_i + \frac{h}{2}(B^T L_{i+1}^T L_{i+1} + B^T L_i^T L_i), \tag{2.6}$$

where $K_0 = 0$ and $L_{-1} = L_0 = C$.

Several remarks may be useful at this point.

Remark 1. The stiffness of the matrix A dictates the choice of stepsize h.

Remark 2. Subtraction of the expression in (2.3) from that in (2.6) yields that the predicted values $K^{(0)}_{i+1}$ and the corrected values K_{i+1} satisfy

$$K_{i+1} - K^{(0)}_{i+1} = \frac{h}{2} B^T (L_{i+1}^T L_{i+1} - 2L_i^T L_i + L_{i-1}^T L_{i-1}) \approx \frac{h}{2} B^T \frac{d^2}{dt^2}(L^T L)$$

and this relationship can be used for stepsize refinement, i.e., to give local bounds depending on stepsize which can be used in error control.

Remark 3. The formula (2.5) can be rewritten as

$$\begin{aligned} L_{i+1} &= L_i\left(I + \frac{h}{2}A_i\right)\left(I - \frac{h}{2}A_i\right)^{-1} \\ &= 2L_i\left(I - \frac{h}{2}A_i\right)^{-1} - L_i, \end{aligned} \tag{2.7}$$

where $A_i \equiv A - BK^{(0)}_{i+1/2}$. Defining $H = I - (h/2)A$ we have that $I - (h/2)A_i = H + (h/2)BK^{(0)}_{i+1/2}$ where $B \in R^{n\times m}$ and $K^{(0)}_{i+1/2} \in R^{m\times n}$. Thus by the Sherman-Morrison-Woodbury formula [29, p. 50] (used frequently when updating an $n \times n$ matrix by rank

m matrices)

$$\left(I-\frac{h}{2}A_i\right)^{-1}=H^{-1}-\frac{h}{2}H^{-1}B\left(I+\frac{h}{2}K_{i+1/2}^{(0)}H^{-1}B\right)^{-1}K_{i+1/2}^{(0)}H^{-1}, \tag{2.8}$$

where $I+(h/2)K_{i+1/2}^{(0)}H^{-1}B\in R^{m\times m}$. The matrix $K_{i+1/2}^{(0)}H^{-1}$ in (2.8) can be computed by

$$\begin{aligned} K_{i+1/2}^{(0)}H^{-1}&=K_iH^{-1}+\frac{h}{2}\left(\frac{3}{2}B^TL_i^TL_iH^{-1}-\frac{1}{2}B^TL_{i-1}^TL_{i-1}H^{-1}\right),\\ K_{i+1}H^{-1}&=K_iH^{-1}+\frac{h}{2}(B^TL_{1+i}^TL_{1+i}H^{-1}+B^TL_i^TL_iH^{-1}). \end{aligned} \tag{2.9}$$

Hence we see from (2.7)-(2.9) that the step (2.5) only involves the operation L_iH^{-1} plus inversion of an $m\times m$ matrix $I+K_{i+1/2}^{(0)}H^{-1}B$. Thus the step (2.5) can be reformulated so that it requires only an $m\times m$ matrix inversion plus matrix-vector multiplications if the LU decomposition of $H=I-(h/2)A$ is computed a priori. This procedure can be most advantageous computationally when m and p are small compared to n.

Remark 4. For some problems we might wish to use a completely implicit scheme in place of (2.3)-(2.6) to enhance stability and reduce sensitivity to step size choice. Then we might consider iterations $K_{i+1}^{(j)}$, $L_{i+1}^{(j)}$ generated by

$$\begin{aligned} L_{i+1}^{(j)}&=L_i+\frac{h}{2}(L_{i+1}^{(j)}+L_i)(A-BK_{i+1/2}^{(j-1)}),\\ K_{i+1}^{(j)}&=K_i+\frac{h}{2}(B^TL_{i+1}^{(j)}L_{i+1}^{(j)}+B^TL_i^TL_i),\\ K_{i+1/2}^{(j)}&=(K_{i+1}^{(j)}+K_i)/2 \end{aligned} \tag{2.10}$$

and thus produce iterates with limits (as $j\to\infty$)K_{i+1}, L_{i+1} satisfying

$$\begin{aligned} L_{i+1}&=L_i+\frac{h}{2}(L_{i+1}+L_i)(A-B(K_{i+1}+K_i)/2),\\ K_{i+1}&=K_i+\frac{h}{2}(B^TL_{i+1}^TL_{i+1}+B^TL_i^TL_i). \end{aligned} \tag{2.11}$$

3. An iterative method for computing the optimal feedback gain K. A widely used iterative method for finding the nonnegative solution of the algebraic Riccati equation (1.2) is Newton's method as modified by Kleinman [23]. We show that this method can be reformulated so that, when combined with a factored form of the well-known Smith method [35], we can compute directly a sequence of iterates K_i for the feedback gain K.

First we recall the Newton iterative algorithm as formulated by Kleinman:

(1) Choose a gain matrix K_0 so that $A-BK_0$ is stability matrix (i.e., Re $\lambda<0$ for all eigenvalues of $A-BK_0$);

(2) Update K_i by $K_{i+1}=B^TP_i$ where P_i is the solution of the Lyapunov equation

$$(A-BK_i)^TP_i+P_i(A-BK_i)+K_i^TK_i+C^TC=0.$$

It has been shown in [23] that $0\leqq P_{i+1}\leqq P_i$ for any i, and $P=\lim P_i$ where the convergence is quadratic. This algorithm can be viewed as an iterative method for the gain K, i.e., $K=\lim K_i$ where $K_{i+1}=F(K_i)$ with $F(K)=B^TP$ and P is the solution of the Lyapunov equation

$$(A-BK)^TP+P(A-BK)+K^TK+C^TC=0. \tag{3.1}$$

Thus, in order to calculate $F(K)$ we must solve the Lyapunov equation (3.1) for the symmetric matrix P. However, we can form an alternative version that allows us to directly calculate $F(K)$ using the Smith method for a Lyapunov equation in X of the form

$$S^TX+XS+D^TD=0, \tag{3.2}$$

where $S\in R^{n\times n}$ is a stability matrix and $D\in R^{p\times n}$.

To this end, we replace step (2) in the Newton–Kleinman method by the following:

(2′) For $i\geqq 1$, update K_i by $K_{i+1}=K_i-B^TZ_i$ where $Z_i=P_{i-1}-P_i$ is the solution of the Lyapunov equation

$$(A-BK_i)^TZ_i+Z_i(A-BK_i)+D_i^TD_i=0 \tag{3.3}$$

with $D_i\equiv K_i-K_{i-1}$.

The method with (2′) offers several advantages over that using (2). The Lyapunov equation in (2′) has fewer inhomogeneous terms than does the one in (2) and the term D_i has rank m which depends only on the number of inputs (controls) to the system. In the proposed modified Smith method described below, we are able to compute directly the $m\times n$ update matrix $J^i=B^TZ_i$ without computing Z_i (see (3.12) below). Since $Z_i\to 0$ as $i\to\infty$ is expected, choosing the start-up value $J_0^i=0$ in the factored Smith algorithm (where $J^i=B^TZ_i$ is computed as the limit as $k\to\infty$ of a sequence J_k^i) is a natural as well as convenient choice.

Note that the step (2′) requires that we have $Z_1=P_0-P_1$ in hand and hence we must start this procedure with P_0, P_1 (and K_0, K_1) given whereas (2) requires only that we start with K_0 given. Then K_1 is computed by $K_1=B^TP_0$ with P_0 the solution of

$$(A-BK_0)^TP_0+P_0(A-BK_0)+K_0^TK_0+C^TC=0.$$

Since our Smith algorithm below is formulated to solve Lyapunov equations of the form (3.2), we can, to maintain this form, initially solve the equation twice. That is, if we solve for $\tilde{Z}_0$ the solution of

$$(A-BK_0)^T\tilde{Z}_0+\tilde{Z}_0(A-BK_0)+K_0^TK_0=0 \tag{3.4}$$

and for $\hat{Z}_0$ the solution of

$$(A-BK_0)^T\hat{Z}_0+\hat{Z}_0(A-BK_0)+C^TC=0, \tag{3.5}$$

then we can obtain K_1 by $K_1=B^T\tilde{Z}_0+B^TZ_0$. Since the Smith method as formulated here actually returns B^TX where X is the solution to (3.2), we thus will use this Smith algorithm twice (with $S=A-BK_0$), once with $D^TD=K_0^TK_0$, once with $D^TD=C^TC$, and then simply add the solutions to obtain K_1.

We turn next to the desired factored form of the Smith method as applied to equation (3.2). Let X_0 be an arbitrary $n\times n$ symmetric matrix and let a sequence $\{X_k\}$ of $n\times n$ symmetric matrices be generated by

$$X_{k+1}=U_r^TX_kU_r+Y_r, \tag{3.6}$$

where r is a positive constant (the Smith stepsize) and

$$U_r=(I-rS)^{-1}(I+rS), \tag{3.7}$$

$$Y_r=2r(I-rS^T)^{-1}D^TD(I-rS)^{-1}. \tag{3.8}$$

Then we can argue [33] that $\{X_k\}$ converges to X, the solution of (3.2). The method and its analysis is based on the observation that for any positive constant r, the equation (3.2) is equivalent to

$$X = U_r^T X U_r + Y_r$$

which can be used to define a contraction map in the obvious manner [33].

We modify this standard formulation of the Smith method to suit our particular needs here (computing $B^T X$ instead of X). From (3.6) we have

$$X_{k+1} - X_k = U_r^T (X_k - X_{k-1}) U_r, \ k \geqq 1.$$

Hence, if $X_k - X_{k-1} = M_k^T M_k$ (i.e., if we have a factorable difference), then

$$X_{k+1} - X_k = U_r^T M_k^T M_k U_r = (M_k U_r)^T (M_k U_r).$$

If the start-up value X_0 is zero, then we can write

$$X_1 - X_0 = 2r M_1^T M_1, \qquad M_1 \equiv D(I - rS)^{-1}. \tag{3.9}$$

By induction on k, (3.6) is then equivalent to

$$M_{k+1} = M_k U_r, \tag{3.10}$$

$$X_{k+1} = X_k + 2r M_{k+1}^T M_{k+1}. \tag{3.11}$$

In this manner $B^T X$ can be obtained as the limit of $J_k = B^T X_k$ where J_k is generated by

$$J_{k+1} = J_k + 2r B^T M_{k+1}^T M_{k+1}.$$

Thus, the update step (2′) is carried out by the following Smith algorithm:

(3.12) (i) Set $S_i = A - BK_i$ and $D = K_i - K_{i-1}$;

(3.12) (ii) Choose a positive constant r and form U_r and M_1 by (3.7) and (3.9) with $S = S_i$; put $J_0 = 0$ and $J_1 = J_0 + 2r B^T M_1^T M_1$;

(3.12) (iii) Iterate for $k = 1, 2, \cdots$, on

$$M_{k+1} = M_k U_r,$$

$$J_{k+1} = J_k + 2r B^T M_{k+1}^T M_{k+1}.$$

In summary, we have described in this section a Newton–Kleinman scheme combined with the Smith method for the resulting Lyapunov equation at each step in the Newton–Kleinman. We have reformulated the Newton–Kleinman iteration and factored the Smith algorithm so as to result in algebraic savings in computing directly the gain estimates K_i.

4. The Smith method and variable stepsize. As is well known, the rate of convergence in the Smith method discussed in the last section depends upon the choice of the acceleration or step parameter. (See [33, p. 291–297] for several discussions. Note that our parameter r is the negative reciprocal of the parameter in Russell's discussions.) To increase speed in convergence, we may employ the accelerated Smith method [33], [35] which can yield quadratic convergence as compared to the linear convergence obtained with (3.6). However, unlike (3.6), the accelerated Smith method is not self-correcting [33] and here we propose to speed up convergence in an alternative way which has proved both reliable and efficient in some of our numerical tests. Specifically, we propose to use a succession of acceleration parameter values r_i (much in the spirit of other well-known iterative methods such as alternating directions [15], [28]) to accelerate convergence in the basic Smith method. Our formulation of this "variable stepsize" Smith method is based upon the observation that for fixed $r>0$ and $k\geqq 1$, the Smith algorithm can be written as

$$S^TX_k+X_kS+D^TD=E_k,\qquad E_k\equiv(I+rS)^TM_k^TM_k(I+rS), \tag{4.1}$$

where M_k is defined by (3.9) and (3.10). To see this, we note that from (3.6) we have

$$(I-rS)^TX_k(I-rS)=(I+rS)^TX_{k-1}(I+rS)+2rD^TD$$

or

$$\begin{aligned}(I+rS^T)X_k(I+rS)-(I-rS)^TX_k(I-rS)+2rD^TD\\ =(I+rS)^T(X_k-X_{k-1})(I+rS).\end{aligned}$$

Hence, from (3.11) we obtain

$$2r(S^TX_k+X_kS+D^TD)=2r(I+rS)^TM_k^TM_k(I+rS),$$

which implies (4.1). Moreover, from (3.9) and (3.10) we have

$$E_k=(U_r^T)^kD^TD(U_r)^k,\qquad k\geqq 1. \tag{4.2}$$

Thus, if we use the iteration (3.6) with acceleration parameter r_1 for k_1 iterates, we obtain an iterate $X^{(1)}$ and equation error $E^{(1)}$ satisfying

$$S^TX^{(1)}+X^{(1)}S+D^TD=E^{(1)},\qquad E^{(1)}=(U_{r_1}^T)^{k_1}D^TD(U_{r_1})^{k_1}. \tag{4.3}$$

Let us define the difference $\Sigma^{(1)}=X-X^{(1)}$ where X is the sought-after solution of (3.2). Then it is readily seen that $\Sigma^{(1)}$ satisfies a Lyapunov equation similar to that of (3.2):

$$S^T\Sigma+\Sigma S+E^{(1)}=0. \tag{4.4}$$

If we next apply the iteration (3.6) k_2 times with acceleration parameter r_2 to the residual equation (4.4) we obtain

$$S^TX^{(2)}+X^{(2)}S+E^{(1)}=E^{(2)}, \tag{4.5}$$

where $X^{(2)}$ is the final iterate using r_2 and the equation error $E^{(2)}$ is given by

$$E^{(2)}=(U_{r_2}^T)^{k_2}E^{(1)}(U_{r_2})^{k_2}.$$

If we proceed to define the difference $\Sigma^{(2)}=X-(X^{(1)}+X^{(2)})$, then from (4.4) and (4.5) we see that $\Sigma^{(2)}$ satisfies a Lyapunov equation

$$S^T\Sigma+\Sigma S+E^{(2)}=0.$$

We continue this procedure, using a sequence of acceleration values $\{r_i\}$ along with corresponding iteration counts $\{k_i\}$ to produce a sequence $\{X^{(i)}\}$ of nonnegative, symmetric matrices. For $i\geqq 1$, we have

$$S^TX^{(i)}+X^{(i)}S+E^{(i-1)}=E^{(i)}, \tag{4.6}$$

$$E^{(i)}=(U_{r_i}^T)^{k_i}E^{(i-1)}(U_{r_i})^{k_i},\qquad E^{(0)}\equiv D^TD. \tag{4.7}$$

Thus, if $\bar{X}_j\equiv\sum_{i=1}^{j}X^{(i)}$, then $\bar{X}_j$ satisfies

$$S^T\bar{X}_j+\bar{X}_jS+D^TD=E^{(j)} \tag{4.8}$$

and hence $\bar{X}_j\leqq X$ and $\bar{X}_{j-1}\leqq\bar{X}_j$, $j\geqq 1$.

Using arguments similar to those in [33, p. 291-294] we can show that for $0<\underline{r}\leqq r\leqq R$, with $\underline{r}$, R positive constants, there exists a constant ω, $0<\omega<1$, depending only on $\underline{r}$ and R, such that for $\omega<\rho<1$,

$$|U_r^k|\leqq M(\rho)\rho^k,\qquad k=0,1,\cdots,$$

where $M(\rho)$ is independent of r. Thus, if $\underline{r} \leqq r_i \leqq R$, then for any $0 < \varepsilon < 1$, there exists an integer $k(\varepsilon)$ such that for $k_i \geqq k(\varepsilon)$, $i \geqq 1$,

$$|U_{r_i}^{k_i}| \leqq 1 - \varepsilon. \tag{4.9}$$

Hence, using (4.7) we have

$$|E^{(j)}| \leqq (1-\varepsilon)^{2j} |D|^2$$

so that $E^{(j)} \to 0$ as $j \to \infty$ and therefore $\bar{X}_j \to X$ as $j \to \infty$.

For the hybrid method proposed in this paper, we have combined the variable stepsize method just outlined with the reformulated Smith method of (3.12). We then obtain the following algorithm for solving for the feedback gain K_i:

ALGORITHM 4.10. Set $S_i = A - BK_i$, $D_1 = K_i - K_{i-1}$ and $J_0 = 0$. For given acceleration parameters $r_1, r_2, \cdots$, and iteration count indices $k_1, k_2, \cdots$, we iterate on $j = 1, 2, \cdots$, in the following steps:

(4.10a) Compute $U_{r_j} = (I - r_j S_i)^{-1}(I + r_j S_i)$ and

$$M_1 = D_j (I - r_j S_i)^{-1},$$
$$J_1 = J_0 + 2 r_j B^T M_1^T M_1.$$

(4.10b) Iterate for $k = 1, 2, \cdots, k_j - 1$ in

$$M_{k+1} = M_k U_{r_j}$$
$$J_{k+1} = J_k + 2 r_j B^T M_{k+1}^T M_{k+1}.$$

(4.10c) Compute $D_{j+1} = M_{k_j}(I + r_j S_i)$, set $J_0 = J_{k_j}$ and return to (4.10a) with $j = j+1$.

The steps in Algorithm (4.10) are to be repeated, i.e., iteration through $r_1, r_2, \cdots$, until some convergence criterion is met. In performing the steps in (4.10b), we can use the Sherman-Morrison-Woodbury formula in a procedure such as that outlined in (2.7)-(2.9) in § 2.

Remarks. (1) The convergence of our hybrid method is guaranteed if (i) $A - BK_0$ is stable where K_0 is obtained through a numerical integration of the Chandrasekhar equation by Algorithm (2.3)-(2.9), and (ii) the iteration counts k_i's in the variable stepsize algorithm Algorithm (4.10) are chosen sufficiently large. Condition (i) can be met if the stepsize h is chosen sufficiently small and the integration is carried out on a sufficiently large interval $[-t_f, 0]$ so that the norm of L at $-t_f$ is smaller than a given small constant, say $\varepsilon > 0$.

(2) The efficiency of the method may rely heavily on a choice of the acceleration parameters $r_1, r_2, \cdots, r_l$. If we follow the guide provided by ADI methods (see [15, p. 37]), we might choose a set of values r_j to be used in some cyclic order. The best choices of values for the r_j often depend on the eigenvalues of $S_i = A - BK_i$. For example, consider the case where S_i has only real eigenvalues λ_j, each with multiplicity m_j, $j = 1, 2, \cdots, m$. Then a choice of $r_j = -1/\lambda_j$ and $k_j = m_j$ in the algorithm produces convergence in a finite number (m) of steps. That is, this choice yields $E^{(m)} = 0$ in (4.8). Of course, the complete eigenstructure of S_i will not be known (nor do we suggest that any sophisticated analysis along these lines be included with each use of Algorithm (4.10) to obtain the gains K_i).

For the case where the problem is obtained from an approximation of an LQR-problem for systems governed by PDE's, a good approximation method preserves the property of solutions to the orignal LQR-problem; e.g., the distribution of closed-loop eigenvalues. Thus, depending on the type of PDE's we consider, we can have an a priori knowledge of bounds on the closed loop eigenvalues for the approximating system. For example, in Example 2 of § 5, we consider a boundary control problem for the one-dimensional heat equation. In this case, we know that the closed-loop eigenvalues are contained in a sector $S=\{\lambda \in C: |\arg(\lambda-\rho)| \leqq (\pi/2)+\theta, \rho \geqq 0 \text{ and } \theta>0\}$. Thus we can choose the parameters r_i systematically as shown in Table 5.3.

(3) For the full matrix case, the following are rough operation counts. For the Chandrasekhar step, we have

$$\frac{n^3}{3}+\frac{t_f}{h}(pn^2+4mpn),$$

where t_f is the terminal time and h is the stepsize. For the Newton–Kleinman step with variable stepsize Smith, we find

$$\frac{l}{3}n^3+c_1(k_1+k_2+\cdots+k_l)[pn^2+4mpn+c_2(mn^2+4m^2n)],$$

where l is the number of acceleration parameters r_i, c_1 is the number of cycles of Algorithm (4.10), and c_2 is the number of Newton iterations. When A is sparse, the operation counts $n^3/3$ for the LU-decomposition of $(I-rA)$ and pn^2 for computing $L(I-rA)^{-1}$ can be significantly reduced using a Gauss-elimination procedure for the matrix A.

In closing this section, we note that the analogies of our variable step Smith method with the ADI methods used to solve partial differential equations can be made a little more precise. Briefly, in ADI splitting methods [28, p. 146–148], we attempt to solve a discretization of the evolution equation

$$\dot{\Phi}=A\Phi+f$$

when $A \geqq 0$ can be written $A=A_1+A_2$ with $A_i \geqq 0$ (for example, factored into components corresponding to spatial discretizations in the x and y directions, respectively, for an equation in a two-dimensional spatial domain). This can be shown [28, p. 150] to lead to an iterative scheme

$$\left(I+\frac{h}{2}A_1\right)\left(I+\frac{h}{2}A_2\right)\Phi^{j+1}=\left(I-\frac{h}{2}A_1\right)\left(I-\frac{h}{2}A_2\right)\Phi^j+hf^j, \tag{4.11}$$

where the index j is related to time stepping. On the other hand, if we consider the Smith method (3.6)-(3.8) for

$$S^TX+XS=F$$

and choose $r=h/2$, we obtain the iteration

$$\left(I-\frac{h}{2}S^T\right)X^{j+1}\left(I-\frac{h}{2}S\right)=\left(I+\frac{h}{2}S^T\right)X^j\left(I+\frac{h}{2}S\right)+hF. \tag{4.12}$$

In these iterations we may identify the $n \times n$ matrix $X=[x_1,\cdots,x_n]$, $x_i \in R^n$ and the n^2 vector $\Phi=\text{column}[x_1,\cdots,x_n]$. If we then identify $A_1\Phi$ with $-S^TX$ (i.e., $A_1=-I\otimes S^T$) and $A_2\Phi$ with $-XS$ (i.e., $A_2=-S\otimes I$), we can immediately see the equivalence between (4.11) and (4.12).

5. Summary remarks and numerical examples. In the preceding sections we have presented an algorithm that offers some definite advantages in computing directly the feedback gains K for high dimensional LQR problems such as those arising in approximating partial or delay differential equation control problems. As we will see with several numerical examples in this section, it can substantially outperform standard eigenvector methods on such problems. As we have pointed out, a fundamental algebraic operation (in both the Chandrasekhar update (2.6), (2.7) and in the reformulated Smith methods (4.10b)) involves computation of

$$L(I-r(A-BK))^{-1}, \tag{5.1}$$

where L and K are $p\times n$ and $m\times n$ matrices, respectively. Our algorithm uses the Sherman–Morrison–Woodbury formula which can provide significant computational savings when m and p are small compared to n. For systems involving sparse matrices A (a frequent occurrence in many approximation schemes), the needed calculations can be carried out quite efficiently.

We further note that that Chandrasehkar and Newton–Kleinman–Smith components as formulated in our algorithm lead to ready estimates between the true gain K and the iterates K_i in terms of equation errors in the steps being performed.

One component (the variable step Smith) of the algorithm is most effectively carried out if we possess some a priori knowledge of bounds on the closed loop eigenvalues. If the closed loop eigenvalues lie close to the imaginary axis, then convergence in the Smith method can be very slow. Eigen or Schur vector methods [27], [30] are less sensitive in this regard. For low order systems, the Schur vector approach is more reliable and less expensive computationally than our algorithm. Our hybrid algorithm depends critically on a number of choices (e.g., stopping criteria in the Newton–Kleinman and Smith components, stepsize sequence $\{r_j\}$ and iteration count sequence $\{k_j\}$ in the variable step component) to be made by the user, and the "best" choices are heavily problem dependent. Hence we can expect our hybrid algorithm to require more experimentation and fine tuning than other more standard methods. However, as we will demonstrate with examples, for the case where n is large compared to m and p, it can offer considerable computational savings with no loss in accuracy over the methods mentioned above.

We have tested (and are continuing our efforts in this direction) our hybrid algorithm on several numerical examples. We will report on just two of these here to illustrate our findings.

In Example 1, our algorithm is tested for the ill-posed LQR-problem reported in [27].

In Example 2, we will demonstrate various aspects of the proposed algorithm, using the approximate LQR-problems via Galerkin approximation of a boundary control problem for the one-dimensional heat equation. All our computations were carried out in double precision on an IBM 3081 at Brown University. We gratefully acknowledge the assistance of Yun Wang in our carrying out of the extensive computational studies reported for the boundary flux control in the diffusion equation problem of Example 2 below.

Example 1. As one of our examples, we considered an example (Example 6 of [27]) which Laub used to test his Schur based methods. The system is the n-dimensional system of (1.1) with

$$A=\begin{bmatrix}0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & & 0\\ \vdots & & & & \\ \vdots & & & 0 & 1\\ 0 & \cdots & & \cdots & 0\end{bmatrix}\qquad B=\begin{bmatrix}0\\ \vdots\\ 0\\ 1\end{bmatrix}\qquad C=[1\quad 0\quad \cdots\quad 0],$$

which leads to an ill-conditioned Riccati equation. This problem corresponds to one in which n integrators are connected in a series with a feedback controller to be applied to the nth integrator in order to stabilize the system. Only deviations of x_1 from the origin are penalized in the cost functional. The true optimal gain is an n vector $K=(\bar{K}^1, \cdots, \bar{K}^n)$ and for this example we can argue that $\bar{K}^1=1$. In [27], Laub used his Schur techniques to study this example and reported difficulties with loss of accuracy at a relatively low value of n, $n=21$. We carried out runs with our hybrid algorithm and obtained quite favorable performance. Some of our findings included the following:

(a) For $n=40$, we used the Chandrasekhar component to integrate to $t_f=100$ and produce an initial estimate $K_0^1=.99041$, which, when used in the Newton-Kleinman (fixed step size $r=.5$ in the Smith), produced the estimate $K_5^1=1.0$—in a total of 2.93 seconds of CPU time. When we used a cruder solution in the Chandrasekhar component ($t_f=200$ but with step size twice that in the first run) to produce $K_0^1=.9394$, followed by the $N-K$ ($r_1=.5$, $r_2=1.0$ in the variable step Smith) we obtained $\mathrm{K}_6^1=1.0$—all in 2.39 seconds.

(b) For $n=50$, we produced $K_0^1=.9224745$ at $t_f=220$ and after the N-K-Smith (fixed step $r=.5$ in the Smith) obtained $K_5^1=1.0000000003820$ in a total of 4.44 CPU seconds. For the same runs with variable step ($r_1=.5$, $r_2=.7$) Smith we obtained a K_5^1 as above with 3820 replaced by 3817 in a total of 4.31 seconds.

(c) We compared runs with the Chandrasekhar component only against the Potter method for $n=10$, 21, 40. Obtaining essentially the same estimates for $n=10$ and 21 (at $n=40$, the Potter degenerates numerically to produce useless estimates) we had CPU times of $\mathrm{CH}_{n=10}=.753$ seconds, $\mathrm{POTT}_{n=10}=.188$ seconds, $\mathrm{CH}_{n=21}=1.52$ seconds, $\mathrm{POTT}_{n=21}=1.22$ seconds, $\mathrm{CH}_{n=40}=4.35$ seconds, $\mathrm{POTT}_{n=40}=6.81$ seconds.

We also implemented the Laub-Schur method using the HQR3 and EXCHNG routines of Stewart [37]. Our implementation was exactly the same as suggested in [27] except that the stable and unstable blocks are separated by their absolute values along the diagonal. For the two examples we considered here, we did not observe any significant difference between Potter's method and Laub's method, either in performance or CPU times.

We found for this example that the eigenvector methods are best for small n, but as n grows, the Chandrasekhar alone, and, even more so, the hybrid method will out perform the Eigen-Schur methods in both accuracy and CPU times. A more striking demonstration of this behavior will be given in the next example.

Example 2. We consider the following linear quadratic regulator problem: minimize the cost functional

$$J(u)=\int_0^\infty (|Cz(t)|^2+|u(t)|^2)\,dt \tag{5.2}$$

subject to the partial differential equation

$$\begin{aligned}\frac{\partial}{\partial t}z(t,x)&=\frac{\partial^2}{\partial x^2}z(t,x), \qquad x\in(0,1),\\ z(0,x)&=\Phi(x)\end{aligned} \tag{5.3}$$

with boundary conditions

$$\frac{\partial}{\partial x}z(t,0)=u(t) \quad \text{and} \quad \frac{\partial}{\partial x}z(t,1)=0 \tag{5.4}$$

and

$$Cz(t)=\int_0^1 c(x)z(t,x)\,dx,$$

where $c(\cdot)$ is square integrable on $[0,1]$.

We can discretize or approximate (5.3)-(5.4) using the standard Galerkin method [2]; i.e., the approximating solution $z^N(t,x)$ to (5.3)-(5.4) is given by

$$z^N(t,x)=\sum_{i=0}^{N} w_i(t)l_i(x), \qquad w_i(t)\in R^1, \tag{5.5}$$

where $l_i=l_i^N$ is the first-order spline defined by

$$l_i^N(x)=\begin{cases} N\left(x-\dfrac{i-1}{N}\right), & \dfrac{(i-1)}{N}\leqq x\leqq \dfrac{i}{N}\\ N\left(\dfrac{i+1}{N}-x\right), & \dfrac{i}{N}\leqq x\leqq \dfrac{(i+1)}{N}\\ 0 \quad \text{otherwise} & \end{cases}$$

and $z^N(t,x)$ satisfies

$$\int_0^1 \frac{\partial}{\partial t} z^N(t,x)\psi^N(x)\,dx=-\int_0^1 \frac{\partial}{\partial x} z^N \frac{\partial}{\partial x}\psi^N\,dx-u(t)\psi^N(0) \tag{5.6}$$

$$\text{for all } \psi^N\in Z^N=\text{span}\,\{l_0^N, l_1^N,\cdots,l_N^N\}.$$

Then, (5.6) leads to the nth-order $(n=N+1)$ ordinary differential equation for $w^N=\text{col}\,(w_0,\cdots,w_N)$;

$$Q^N\dot{w}^N(t)=-H^Nw^N(t)-B^Nu(t), \tag{5.7}$$

where

$$Q^N=\frac{1}{N}\begin{bmatrix} \frac13 & \frac16 & 0 & \cdots & & 0\\ \frac16 & \frac23 & \frac16 & & & \vdots\\ 0 & & & & & \vdots\\ \vdots & & & & & 0\\ \vdots & & & \frac16 & \frac23 & \frac16\\ 0 & \cdots & & 0 & \frac16 & \frac13 \end{bmatrix} \quad \text{with } Q_{ij}^N=\int_0^1 l_il_j\,dx,$$

$$H^N=N\begin{bmatrix} 1 & -1 & 0 & \cdots & & 0\\ -1 & 2 & -1 & & & \vdots\\ 0 & & & & & \vdots\\ \vdots & & & & & 0\\ \vdots & & & -1 & 2 & -1\\ 0 & \cdots & & 0 & -1 & 1 \end{bmatrix} \quad \text{with } H_{ij}^N=\int_0^1 \frac{d}{dx}l_i\frac{d}{dx}l_j\,dx,$$

and

$$B^N=\text{col}\,(1 \quad 0 \quad \cdots \quad 0).$$

For computational convenience, we change coordinates (for fixed N) in the system (5.7) by $x=Q^Nw^N$ to obtain the approximate system

$$\dot{x}=-H^N(Q^N)^{-1}x-B^Nu.$$

Thus, in (1.1) we have $A=-H^N(Q^N)^{-1}$, $B=-B^N$ and $C=C^N(Q^N)^{-1}$ where C^N is the vector with components $c_i^N=\int_0^1 c(x)l_i(x)\,dx$, $0\leqq i\leqq N$.

For the problem in this example, the approximating optimal feedback operator K^N is given [2] by

$$K^N z = \int_0^1 k^N(x) z(x)\, dx,$$

where $k^N(x) = \sum_{i=1}^N k_i l_i(x)$ and $K = (k_0, \cdots, k_N)$ is the optimal feedback solution in the problem for (1.1) with A, B, C chosen as indicated above. We note in this case that for any $N \geqq 1$, A has only one unstable eigenvalue (zero), (A, B) is stabilizable, and (A, C) is detectable.

For the special case when $c(x) = 1$, we find $C = (1, \cdots, 1) \in R^{1\times(N+1)}$ and hence $CA = 0$. It is thus easy to see that the desired solution $(K(t), L(t))$ to the Chandrasekhar system (1.3) is given by

$$K(t) = k(t)C, \qquad L(t) = l(t)C,$$

where k, l are scalar functions satisfying

$$\dot{k} = l^2, \qquad k(0) = 0,$$

$$\dot{l} = -lk, \qquad l(0) = 1.$$

Therefore we find $\dot{k}k + \dot{l}l = 0$ so that $k^2(t) + l^2(t) = 1$. We thus find in this case that $k(t) \to 1$ as $t \to -\infty$ and hence $K = \lim K(t) = C$. For this case, the Chandrasekhar system for the infinite dimensional LQR problem (5.2)-(5.4) can also be analyzed [17], [36] and exactly the same argument as above shows the optimal feedback gain operator is given by

$$Kz = \int_0^1 1 \cdot z(x)\, dx.$$

These analytic solutions can be used to test software packages and approximation schemes before more interesting, analytically intractable examples are considered.

Remark. The form (5.7) of system equations appears frequently in applications. Thus the critical computational factor (5.1) can be modified so that we can avoid computing A. For example, in this case it has the form

$$L(I - r(-HQ^{-1} - BK))^{-1} = LQ(Q + rH + rBKQ)^{-1}, \tag{5.8}$$

where $Q + rH$ is a symmetric, tridiagonal, positive matrix. Thus we can readily use the Cholesky decomposition algorithm for computing $LQ(Q + rH)^{-1}$ and combine this with the Sherman-Morrison-Woodbury formula (see Remark 3 of § 2) to efficiently compute the critical expression (5.8).

We carried out extensive computations for this example with $c(x) = 1 + x$. We compared our hybrid method to the Potter algorithm and to the use of the Chandrasekhar system alone. We also used the Laub-Schur method on this example but, as in Example 1 above, found essentially no significant difference between Potter's method and the Laub-Schur, either in performance or CPU times. (Analysis and computational experience indicate that the Potter method and the Laub-Schur method are both $O(N^3)$ with the latter method about twice as fast as the Potter method.) We required, whenever feasible, the same level of accuracy in computation of feedback gains and compared relative CPU times.

In studying our hybrid scheme, we tested numerous sets of Smith acceleration parameters $\{r_j\}$, $\{k_j\}$, stopping times t_f in the Chandrasekhar component and error stopping criteria in both the Chandrasekhar and Newton-Kleinman-Smith components. We summarize some of our findings to date.

In Table 5.1 we present comparative CPU times for the hybrid scheme versus Potter as we increase N. Recall that the corresponding finite-dimensional approximation scheme has system with dimension $n = N+1$. In all of the runs reported in Table 5.1, the feedback gains for the hybrid and Potter calculations agreed to nine decimal places so both schemes provided accurate solutions. In these runs, the hybrid scheme calculations used $t_f = 2.2$ (corresponding to $h = .1$) with $|L(-t_f)| \simeq 10^{-3}$ in the Chandrasekhar component. The Newton-Kleinman component converged after four iterations (i.e., at K_4) and we used acceleration steps $r_1 = 1$, $r_2 = 10^{-1}$, $r_3 = 10^{-3}$, $r_4 = 10^{-5}$. Each Smith iteration was allowed a maximum of $k_j = 50$ per value of r_j although in most cases the iteration satisfied a convergence criterion before this maximum was attained. Careful consideration of Table 5.1 reveals that the hybrid scheme is clearly $O(N)$ while the Potter is $O(N^3)$; both rates are to be expected from our earlier observations about the methods. Note that at $N = 80$ the hybrid scheme is more than 25 times faster than the Potter scheme (with comparable accuracy, of course).

We also ran the hybrid scheme with $N = 80$ and a number of different fixed acceleration values r in the Smith component. The same Chandrasekhar component parameters as reported above were used. Table 5.2 contains relative CPU times as well as an indication of the N-K iterate for which convergence was achieved.

In Table 5.3 we list some CPU times when different sets of acceleration parameters $\{r_j\}$ were used. Again these runs were for $N = 80$ with the same Chandrasekhar solution as above. All of the converged Newton-Kleinman iterates were after six steps (i.e., K_6).

Finally, we made runs (for $N = 80$) to find the best results that the Chandrasekhar algorithm alone (i.e., accurate integration until $K(t) \to K$, $L(t) \to 0$) could produce. The

TABLE 5.1

N	Hybrid (CPU sec.)	Potter (CPU sec.)
10	.17	.14
20	.31	.81
30	.56	2.45
40	.74	5.49
50	.91	10.71
60	1.09	18.09
70	1.26	27.97
80	1.43	41.56
100	1.76	
120	2.10	
140	2.45	
160	2.80	

TABLE 5.2

r	CPU (sec.)	Converged N-K Gain
5	5.10	K_6
1	6.52	K_4
10^{-1}	6.27	K_4
10^{-2}	7.40	K_4
10^{-3}	10.10	K_4
10^{-4}	12.06	K_5
10^{-5}	10.07	K_4

TABLE 5.3

r	CPU (sec.)
$(10^{-1}, 10^{-2})$	6.25
$(1, 10^{-1}, 10^{-2})$	4.88
$(1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5})$	1.98
$(1, 10^{-1}, 10^{-3}, 10^{-5}, 10^{-6}, 10^{-7})$	1.61

best results we were able to achieve yielded an accurate value of K for $K(-t_f)$ with $t_f = 3.22$ with $|L(t_f)| \simeq 10^{-6}$ obtained in 5.85 CPU seconds.

Based on our computational findings for the above two examples and our experience with several other examples for infinite-dimensional systems (e.g., beams with tip bodies, etc.), we are quite confident that the hybrid scheme we propose in this paper can be profitably used with a number of large scale LQR problems. We are currently developing a rather general software package that implements the hybrid scheme in a manner so that a broad range of problems can be treated in the context of the ideas presented here.

REFERENCES

[1] H. T. BANKS AND J. A. BURNS, *Hereditary control problems: numerical methods on averaging approximations*, SIAM J. Control Optim., 16 (1978), pp. 169-208.

[2] H. T. BANKS AND K. KUNISCH, *The linear regulator problem for parabolic systems*, SIAM J. Control Optim., 22 (1984), pp. 684-698.

[3] H. T. BANKS AND I. G. ROSEN, *Computational methods for the identification of spatially varying stiffness and damping in beams*, LCDS Technical Report 86-39, Brown University, August 1986; Control: Theory and Advanced Technology 3 (1987), pp. 1-32.

[4] ———, *Parameter estimation techniques for distributed systems*, in *Control and Estimation in Distributed Parameter systems*, SIAM Frontiers in Applied Mathematics, to appear.

[5] H. T. BANKS, I. G. 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. Comp., 5 (1984), pp. 830-855.

[6] J. A. BURNS AND R. H. FABIANO, *Modeling and approximation for a viscoelastic control problem*, Proc. Conf. Control of Partial Differential Equations, July 7-11, 1986, Vorau, Lecture Notes in Control and Information Sciences 102, 1987, Springer-Verlag, pp. 23-39.

[7] J. A. BURNS, K. ITO, AND R. K. POWERS, *Chandrasekhar equations and computational algorithms for distributed parameter systems*, Proc. 23rd IEEE Conference on Decision and Control., Dec., 1984, Las Vegas, NV.

[8] J. A. BURNS AND R. POWERS, *Factorization and reduction methods for optimal control of hereditary systems*, Math. Appl. Comput., 5 (1987), pp. 203-248.

[9] J. CASTI, *Dynamical Systems and Their Applications: Linear Theory*, Academic Press, New York, 1977.

[10] S. D. CONTE, *Elementary Numerical Analysis*, McGraw Hill, New York, 1965.

[11] J. S. GIBSON, *An analysis of optimal modal regulation: convergence and stability*, SIAM J. Control Optim., 19 (1981), pp. 686-707.

[12] ———, *Linear-quadratic optimal control of hereditary differential systems: infinite dimensional Riccati equations and numerical approximations*, SIAM J. Control Optim., 21 (1983), pp. 95-139.

[13] J. S. GIBSON AND A. ADAMIAN, *Approximation theory for* LQG *optimal control of flexible structure*, SIAM J. Control Optim., 29 (1991), pp. 1-37.

[14] ———, *A comparison of three approxiation schemes for optimal control of a flexible structure*, in Control and Estimation in Distributed Parameter Systems, SIAM Frontiers in Applied Mathematics, to appear.

[15] L. A. HAGEMAN AND D. M. YOUNG, *Applied Iterative Methods*, Academic Press, New York, 1981.

[16] K. ITO, *Legendre-tau approximation for functional differential equations* III. *Eigenvalue approximations and uniform stability*, in Distributed Parameter Systems, F. Kappel, K. Kunisch, W. Schappacher, eds., Lecture Notes in Control and Information Science, 75, Springer-Verlag, Berlin, New York, 1985, pp. 191-212.

[17] K. ITO AND R. K. POWERS, *Chandrasekhar equations for infinite dimensional systems*, ICASE Report 84-67, NASA Langley Research Center, Hampton, VA; SIAM J. Control Optim., 25 (1987), pp. 596-611.

[18] K. ITO AND R. TEGLAS, *Legendre-tau approximation for functional differential equations* II. *The linear quadratic optimal control problem*, SIAM J. Control Optim., 25 (1987), pp. 1379-1408.

[19] J.-N. JUANG AND L. G. HORTA, *Effects of atmosphere on slewing control of a flexible structure*, AIAA/ASME/ASCE/AHS 27th Structures, Structural Dynamics, and Materials Conf., May 19-21, 1986, San Antonio, TX, Paper No. 86-1001-CP.

[20] T. KAILATH, *Some Chandrasekhar-type algorithm for quadratic regulators*, Proc. IEEE Conference on Decision and Control, New Orleans, LA, 1972, pp. 219-223.

[21] F. KAPPEL AND G. PROPST, *Approximation of feedback control for delay systems using Legendre polynomials*, Conf. Sem. Mat. Univ. Bari, 201 (1984), pp. 1-36.

[22] F. KAPPEL AND D. SALAMON, *Spline approximation for retarded systems and the Riccati systems*, SIAM J. Control Optim., 25 (1987), pp. 1082-1117.

[23] D. L. KLEINMAN, *On an iterative technique for Riccati equations computations*, IEEE Trans. Automat. Control, AC-13 (1968), pp. 114-115.

[24] K. KUNISCH, *Approximation schemes for the linear quadratic optimal control problem associated with delay equations*, SIAM J. Control Optim., 20 (1982), pp. 506-540.

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

[26] I. LASIECKA AND R. TRIGGIANI, *The regulator problem for parabolic equations with Dirichlet boundary control* II: *Galerkin approximation*, Appl. Math. Optim. 16 (1987), pp. 198-216.

[27] A. J. LAUB, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Control, AC-24 (1979), pp. 913-921.

[28] G. I. MARCHUK, *Methods of Numerical Mathematics*, Springer-Verlag, New York, 1975.

[29] J. ORTEGA AND W. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[30] J. E. POTTER, *Matrix quadratic solutions*, SIAM J. Appl. Math., 14 (1966), pp. 496-501.

[31] R. K. POWERS, *Chandrasekhar equations for distributed parameter systems*, Ph.D. thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1984.

[32] G. PROPST, *Piecewise linear approximation for hereditary control problems*, Technical Report No. 60-1985, Inst. for Math., Univ. Graz, Graz, Austria.

[33] D. L. RUSSELL, *Mathematics of Finite Dimensional Control Systems: Theory and Design*, Lecture Notes in Pure and Applied Math., vol. 43, Marcel Dekker, New York, 1979.

[34] ———, *On mathematical models for the elastic beam with frequency-proportional damping*, in Control and Estimation in Distributed Parameter Systems, H. T. Banks, ed., SIAM Frontiers in Applied Mathematics, Society for Industrial and Applied Mathematics, Philadelphia, PA, to appear.

[35] R. A. SMITH, *Matrix equation $XA+BX=C$*, SIAM J. Appl. Math., 16 (1968), pp. 198-201.

[36] M. SORINE, *Sur le semi-groupe non lineaire associe a l'equation de Riccati*, INRIA Report No. 167, October 1982 Institut National de Recherche en Informatique et en Automatique (INRIA), Le Chesnay, France.

[37] G. W. STEWART, *Algorithm* 506, HQR3 *and* EXCHNG: *Fortran subroutine for calculating and ordering the eigenvalues of a real upper Hessenberg matrix*, ACM Trans. Math. Software, 2 (1976), pp. 275-280.

[38] W. M. WONHAM, *On a matrix Riccati equation of stochastic control*, SIAM J. Control, 6 (1968), pp. 681-699.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 3, pp. 516-534, May 1991

002

ON CONSTRUCTING NONLINEAR OBSERVERS*

ANDREW R. PHELPS†

Abstract. Two algorithmic approaches to constructing nonlinear observers currently exist. Here one of these is improved by finding an explicit solution to the partial differential equations describing the change of state coordinates, thereby avoiding expensive bracket computations. This simplifies the algorithm and has implications for system identification.

Key words. nonlinear observer, observer normal form, canonical form, linearized error dynamics, observer algorithm, Macsyma, coefficient compatibility

AMS(MOS) subject classifications. 93C10

1. Necessary and sufficient conditions for observer normal form. We consider an uncontrolled affine control system

$$\dot{\xi}=f(\xi), \qquad y=h(\xi), \tag{1}$$

where f is a given vector field and we are operating in the neighborhood of a point $\xi^0=\xi(0)$. The problem is to estimate the state $\xi \in \mathbf{R}^n$ based on the measurement $y \in \mathbf{R}^p$. If we move beyond the special case where f is linear, this estimation problem is not very well understood. This paper takes up one of the current approaches to this problem and solves a certain set of differential equations not hitherto known to be solvable. This yields some amelioration of the algorithmic problems for computation of some nonlinear observers.

The *observer normal form* approach to nonlinear observers involves nonlinear transformation of the state and output coordinates to provide for observer design with linearized error dynamics.

To do this, we introduce new observer (state estimator) dynamics

$$\dot{\zeta}=\hat{f}(\zeta).$$

In this expression $\zeta \in \mathbf{R}^n$ is the observer, where $\hat{f}$ "adjusts" f. We consider the error estimate $e=\xi-\zeta$, seeking to dampen the error exponentially as a function of time.

To get the observer, we separate the influence of the inputs and outputs so as to get a system of the form (1). We then employ two particular normal forms, observ*able* and observ*er* forms. Nonlinear observable form, generally speaking, will be attainable under a wide range of conditions, whereas nonlinear observer form is more problematic. Attaining observer form means that we have found a coordinate system for the system (1) which makes it possible to "read off" the dynamics of the observer. This is the practical motivation underlying the use of these normal forms.

In the linear case, these forms are:

$$\begin{array}{ll} \text{Observable form} & \dot{\xi}=A\xi-B\alpha\xi, \quad y=C\xi, \\ \text{Observer form} & \dot{x}=Ax-\alpha Cx, \quad \bar{y}=\gamma Cx. \end{array}$$

For clarity of notation, we use (ξ, y) for generic state coordinates and reserve $(x, \bar{y})$ for observer form state coordinates.

* Received by the editors January 11, 1988; accepted for publication (in revised form) January 9, 1990.

† Department of Mathematics, University of California, Berkeley, California 94720. Present address, Department of Mathematics, University of California, Davis, California 95616. This research was supported in part by Air Force Office of Scientific Research grant 85-0267.

Here, A, B, and C represent Brunovský canonical form [2] matrices in prime form [16]. Thus, we assume we have *observability indices* $l_1 \leqq \cdots \leqq l_j \leqq \cdots \leqq l_p$, with $\sum_{j=1}^{p} l_j = n$. A is an $n \times n$ block diagonal matrix where the jth block is an $l_j \times l_j$ upper diagonal matrix made up of 1's. C is a $p \times n$ block diagonal matrix where the jth block is a $1 \times l_j$ matrix $[1\,0 \cdot \cdot\, 0]$. B is the $n \times p$ "pseudocontrol" matrix corresponding to C—block diagonal with the jth block being a $l_j \times 1$ matrix $[0 \cdots 0\; 1]^T$. For example, in the case where $n=3$ and $p=1$, Brunovský form would give us

$$A=\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad B=\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad C=[1 \quad 0 \quad 0].$$

Also, α and γ are constant matrices of the appropriate dimensions.

In the nonlinear case, locally we can define (nonlinear) observability indices: In a neighborhood of the operating point ξ^0, we have l_j's, $1 \leqq j \leqq p$, as above, such that

$$\dim \operatorname{span}\{L_f^{i-1} h_j(\xi)\colon 1 \leqq i \leqq l_j \text{ and } 1 \leqq j \leqq p\} = n.$$

Using the above, we have nonlinear normal forms

$$\text{(2)} \qquad \text{Observable form} \qquad \dot{\xi} = A\xi - B\alpha(\xi), \quad y = C\xi,$$

$$\text{(3)} \qquad \text{Observer form} \qquad \dot{x} = Ax - \alpha(Cx), \quad \bar{y} = \gamma(Cx).$$

Note that the α of nonlinear observable form is different than that of the corresponding observer form. See [12] for a more detailed exposition.

We can achieve nonlinear observable form generically. Two approaches, involving different sets of conditions, have been developed for putting a system in *nonlinear observer form.* See, for instance, [1], [4], [8]-[10], [12]-[15], [17], and [18]. We describe the computational implications of their algorithms.

For the following, we annotate systems as follows. We write l_j, $1 \leqq j \leqq p$, for the jth observability index, corresponding to the output y_j, and f_j for the (l_j)th derivative of y_j, as it appears in observable form. We label the state coordinates with double indices $j\colon k$, for $1 \leqq j \leqq p$ and $1 \leqq k \leqq l_j$. Thus, $\xi_{j:1} = y_j$ and $\xi_{j:k}$ is its $(k-1)$th derivative. The unit vector in the $\xi_{j:l_j}$ direction is B_j, from the prime B matrix above. We annotate observable form coordinates with a bar when the output $\bar{y}$ is the particular output associated with observer form (see § 2). These state coordinates are given by $\bar{\xi}$; their observable form polynomials are given by $\bar{f}_j$.

The approach adhered to in this paper was developed in [13] and [14]. This algorithm is determined by the conditions required for conversion of a system (1) to observer form (3). These conditions are:

Observable form. Must be able to convert system to observable form (2);

Output coordinate change. Must satisfy differential equation (d.e.) for $y = y(\bar{y})$;

Polynomial degree. Polynomials $f_j(\xi)$ (the entries of $-B\alpha(\xi)$ in observable form (2)) must have degree $\leqq l_j$, for $1 \leqq j \leqq p$;

Bracket vanishing. Brackets of elements in $\{ad_{-f}^{i-1} B_j\colon 1 \leqq i \leqq l_j\}$ must vanish.

Examining this algorithm computationally, we may break it down into the following steps:

Bracket Vanishing Algorithm (Krener-Isidori-Respondek).

- Transform to observable form.
- Verify degree condition.
- Solve differential equation for $\bar{y}(y)$.
- Change to observable form coordinates $\bar{\xi}$ corresponding to $\bar{y}$ output coordinates.

- Compute $\mathrm{ad}^{i}_{-\bar{f}}\, B_j$, $1 \leqq i \leqq l_j - 1$, $1 \leqq j \leqq p$.
- Check bracket vanishing condition.
- Find observer coordinates and injections.
- Compute Jacobian $\partial\xi/\partial x$.
- Find observer.

A distinct approach has been developed in [1], [4], and [15]. It has been developed only for the case where $p=1$ and does not address the problem of changing output coordinates. (It does, however, incorporate the time-varying case.) It calls for the existence of observable form (2), but not its computation, and replaces the last two conditions above with a single requirement, as follows:

Observable form. Must be able to convert control system to observable form;

Output coordinate change. Must satisfy differential equation for $y = y(\bar{y})$;[1]

Bracket spanning. We must have $d(\mathrm{ad}^{l}_{-f}B) \in \mathrm{span}\,\{dh\}$ (expressed in observer coordinates), modulo a slight technical adjustment on the last coordinate.

This algorithm breaks down computationally as follows:

Bracket Spanning Algorithm (Bestle–Zeitz–Li–Tao).

- Check observable form rank condition.
- Verify observable form has single output index.
- Compute $\mathrm{ad}^{k}_{-f}B$, $0 \leqq k \leqq l$.
- Get $d\alpha/dy$ in observer coordinates.
- Get differential expression $\lambda(\xi)$ (not described here), to adjust check for α_l.
- Check bracket spanning condition.
- Recursively compute $x = x(\xi)$.
- Backsolve for $\xi = \xi(x)$.
- Compute observer.

Experience shows that to solve a general system using the bracket vanishing or bracket spanning conditions appears to be somewhat difficult. The $n=2$ case was completely worked out independently by authors using each method, but the solution for dimension three, in the single output case, has been published incorrectly [14], while dimensions four, five, and further get increasingly complicated to compute. We initially tried to implement this computation as a *Macsyma* program. Our results suggested to us a way through this difficulty, which is described below and developed in the succeeding sections.

The particular difficulty of both these algorithms is their requirement of extensive Lie bracket calculations. In § 2, we develop a substitute for the bracket vanishing condtion which obviates the necessity for these calculations.

Coefficient compatibility. Observable form polynomials must evaluate to certain integrals of differential expressions in injection terms (the entries of $\alpha(Cx)$ in (3)).

In § 3, we further develop the coefficient compatibility condition algorithm by finding a direct formula for the integrals in general observable form coordinates. Section 4 is a summary with concluding remarks.

2. Polynomial coefficients in standard coordinates.

We seek to simplify the computations for nonlinear observer form.

[1] This condition improves this algorithm, which in its orginal form neglected the issue of change of output coordinates.

The polynomial degree condition means that the observable form functions $f_j(\xi)$ must be *polynomials* with coefficients being functions in the observed states $\xi_{j:1}$, $1 \leqq j \leqq p$.

The bracket vanishing condition and the bracket spanning condition provide processes for verifying the convertibility of an observable form polynomial into nonlinear observer form. They do not, however, give us a clear and explicit idea of the restrictions on the coefficients that are necessary and sufficient for the existence of this transformation. They give us enough machinery to arrive at a solution recursively, while leaving us subject to a "bewildering" complexity of computation in all but the simplest cases. We must compute iterated brackets and we end up with complicated and unsuggestive expressions. (See, however, the work of Grossman, e.g., [6], for one approach to improving this situation.)

Suppose we try to find a direct solution of the change of state coordinates $x = x(\xi)$. We compute the d.e.'s stemming from the relationship $\bar{y} = \bar{y}(y)$ and the successive Lie differentiations of the output terms.

So doing, we get

$$\begin{aligned}
\xi_1 &= y(\bar{y}) = y(x_1),\\
\xi_2 &= \dot{\xi}_1 = J(x_2 - \alpha_1),\\
\xi_3 &= \dot{\xi}_2 = J\left(x_3 - \alpha_2 - \frac{d\alpha_2}{d\bar{y}}(x_2 - \alpha_1)\right) + \frac{dJ}{d\bar{y}}(x_2 - \alpha_1),\\
f(\xi) &= \dot{\xi}_l = J\left(-\alpha_l - \frac{d\alpha_{l-1}}{d\bar{y}}(x_2 - \alpha_1) + \cdots\right) + \cdots,
\end{aligned} \tag{4}$$

where J is an abbreviation for $dy/d\bar{y}$. Note that the expressions on the right-hand side of (4) exhibit a cascading complexity of terms in higher derivatives of J's, recurring in repeated backsubstitutions of their lower order versions.

However, we might assume that $J = 1$, that y is actually the "correct" output coordinate function $\bar{y}$, given by the output ordinary differential equation (o.d.e.)[2]. In this case, the equations (4) do indeed become more tractable.

The assumption $J = 1$, or $J = I_p$ (the identity matrix) in the general case, means, in effect, that we are using the particular observable form for our system that is given by using the special output $\bar{y}$ gained from the output d.e. and repeated Lie differentiations of it. We call the coordinate system thus derived *standard coordinates*, annotated by $\bar{\xi}$.

In these coordinates, the d.e. expansion (4) becomes

$$\begin{aligned}
\bar{\xi}_1 &= \bar{y} = x_1,\\
\bar{\xi}_2 &= \dot{\bar{\xi}}_1 = x_2 - \alpha_1,\\
\bar{\xi}_3 &= \dot{\bar{\xi}}_2 = x_3 - L_{\bar{f}}\,\alpha_1 - \alpha_2,\\
&\vdots\\
\bar{\xi}_l &= \dot{\bar{\xi}}_{l-1} = x_l - \sum_{i=1}^{l-1} L_{\bar{f}}^{l-i-1}\alpha_i,\\
\bar{f}_1(\bar{\xi}) &= \dot{\bar{\xi}}_l = \sum_{i=1}^{l} L_{\bar{f}}^{l-i}\,\alpha_i.
\end{aligned} \tag{5}$$

[2] In the bracket spanning condition, this assumption is the starting place for the calculation. Here, however, we have reduced it to a *convenience* with which we will eventually dispense (§ 3).

First, we give some notation. We let

$$\xi^{\#} := \prod_{j=1}^{p} \prod_{k=1}^{r_j} \bar{\xi}_{j:i_{j:k}+1}^{e_{j:k}} \tag{6}$$

be a general monomial in $\bar{f}_m(\bar{\xi})$. Let

$$\bar{a}_{m:\cdots}(\bar{y}) := \bar{a}_{m:\cdots\underbrace{|\cdots\underbrace{i_{j:k}+1\cdots i_{j:k}+1}_{e_{j:k}\text{ times}}\cdots|}_{(j)}\cdots}(\bar{y})$$

denote the coefficient of $\xi^{\#}$. Here, $e_{j:k}$ is the *exponent* of this contribution to the term, r_j is the number of contributions by different time-derivatives of $\bar{\xi}_{j:1} = \bar{y}_j$, and $i_{j:k}$ is the *degree*. The $p-1$ *vertical bars* separate the contributions of each set of time derivatives, with "$\|$" showing an empty contribution from the corresponding $\bar{\xi}_{j:k}$'s, for $1 \leqq k \leqq r_j$. (Note that we use "$i_{j:k}+1$" instead of "$i_{j:k}$" because the degree $i_{j:k}$ is one less than the subscript.) We write $e_j = \sum_{k=1}^{r_j} e_{j:k}$, $e = \sum_{j=1}^{p} e_j$. We also need the *total degree* $w = \sum_{j=1}^{p} \sum_{k=1}^{r_j} i_{j:k}\, e_{j:k}$.

Now, it turns out that the differential equations (5) are readily solvable in standard coordinates, with the proof to be established by a simple induction. Here is a low-dimensional example of how these d.e.'s solve.

Example 2.1. Change of state coordinates, single output, dimension three. The differential equation for $\bar{y} = \bar{y}(y)$ is

$$\frac{d\bar{y}}{dy} = \exp\left(-\frac{1}{3}\int_{\nu^0}^{\nu} a_{23}\, d\nu\right).$$

It solves to

$$\bar{y} = \int_{y^0}^{y} \exp\left(-\frac{1}{3}\int_{\nu^0}^{\nu} a_{23}\, d\eta\right) d\nu.$$

To get the change of coordinates $\bar{\xi} = \bar{\xi}(\xi)$, we iteratively substitute our solutions for the first $j-1$ variables $\bar{\xi}_1, \cdots, \bar{\xi}_{j-1}$ in the equation for $\bar{\xi}_j$, etc. We get

$$\begin{aligned} \bar{\xi}_1 &= \int_{y^0}^{y} \exp\left(-\frac{1}{3}\int_{\nu^0}^{\nu} a_{23}\, d\eta\right) d\nu, \\ \bar{\xi}_2 &= \frac{d\bar{y}}{dy}\,\xi_2, \\ \bar{\xi}_3 &= \frac{d\bar{y}}{dy}\left(\xi_3 - \frac{1}{3}\,a_{23}\xi_2^2\right). \end{aligned} \tag{7}$$

The function $\bar{f}_1(\bar{\xi})$ is written as

$$\bar{a}_{23}\bar{\xi}_2\bar{\xi}_3 + \bar{a}_3\bar{\xi}_3 + \bar{a}_{222}\bar{\xi}_2^3 + \bar{a}_{22}\bar{\xi}_2^2 + \bar{a}_2\bar{\xi}_2 + \bar{a}_1, \tag{8}$$

applying the polynomial degree condition. The d.e.'s (5) then become

$$\bar{\xi}_1 = \bar{y} = x_1, \qquad \bar{\xi}_2 = \dot{\bar{\xi}}_1 = x_2 - \alpha_1,$$

$$\bar{\xi}_3 = \dot{\bar{\xi}}_2 = x_3 - \frac{d\alpha_1}{d\bar{y}}\,\bar{\xi}_2 - \alpha_2,$$

$$\bar{f}_1(\bar{\xi}) = \dot{\bar{\xi}}_3 = -\frac{d\alpha_1}{d\bar{y}}\,\bar{\xi}_3 - \frac{d^2\alpha_1}{d\bar{y}^2}\,\bar{\xi}_2^2 - \frac{d\alpha_2}{d\bar{y}}\,\bar{\xi}_2 - \alpha_3.$$

We see that we can read off the $\bar a$ coefficients and we have, in fact,

$$\bar a_1 = -\alpha_3, \quad \bar a_2 = -\frac{d\alpha_2}{d\bar y}, \quad \bar a_{22} = -\frac{d^2\alpha_1}{d\bar y^2},$$

$$\bar a_{222} = 0, \quad \bar a_{23} = 0, \quad \bar a_3 = -\frac{d\alpha_1}{d\bar y}.$$

Plainly, we can backsolve for the α's by integration. □

In general, and in fact whenever the observability indices are all identical, we can read off the solutions to the $\bar a$ polynomial coefficients in terms of the α injection functions. The coefficients are those that arise naturally from tree structures stemming from partitions in graded differential algebras [5]. We get Theorem 2.2.

THEOREM 2.2. *Suppose that all observability indices are equal, i.e., $l_m = l$ for $1 \leqq m \leqq p$. Then the polynomial coefficient $\bar a_{m:\dots}(\bar y)$ is given by*

$$-\left(\frac{w!}{\prod_{j=1}^{p}\prod_{k=1}^{r_j} e_{j:k}!\,i_{j:k}!^{e_{j:k}}}\right)\frac{\partial^e \alpha_{m:l-w}(\bar y)}{\partial \bar y_1^{e_1}\cdots\partial \bar y_p^{e_p}}. \tag{9}$$

Proof. Outline. We use induction on l. We make a counting argument with combinatorics to trace the typical coefficient of a monomial term.

Body of proof. Let m, $1 \leqq m \leqq p$, be fixed. First of all, if $l = 1$, then the system degenerates to $\dot{\bar\xi}_{m:1} = \bar f_m(\bar\xi) = \bar f_m(\bar y)$. Thus, $\bar a_{m:1} = \bar f_m(\bar y) = -\alpha_{m:1}(\bar y)$, where we take $\alpha_{m:1} := -\bar f_m$. But this matches formula (9).

Note that the result of Example 2.1 also corresponds with formula (9). The derivation shown there illustrates the pattern we use for our induction.

As induction hypothesis, we assume that formula (9) holds for $l = \mu$. We target the coefficients in the expansion for $\bar\xi_{m:\mu} - x_{m:\mu}$ in the case $l = \mu + 1$ that are the *source* via Lie differentiation of the coefficients of $\bar f_m$ we seek to evaluate.

But, when we examine the p.d.e. expansion (5), we find the same expression, $-\sum_{i=1}^{\mu} L_{\bar f}^{\mu-i}\alpha_{m:i}$, for $\bar\xi_{m:\mu} - x_{m:\mu}$ in the expansion with $l = \mu + 1$, as we find for $\bar f_m$ in the expansion with $l = \mu$. This means that the induction assumption will enable us to know those "target" coefficients, which, when Lie differentiated, give contributions to the expansion for $\bar f_m$ in the case when the multi-index l is $\mu + 1$. These coefficients evaluate to nothing but the coefficients of the terms of $\bar f_m$ in the case (given by induction assumption) that the multi-index l is μ.

An expression that under Lie differentiation by $\bar f$ can give a term such as (6), with coefficient $\bar a_{m:\dots}$, may have two forms.

Case 1. It may have no increment in the exponent $e_{j:1}$ of $\bar\xi_{j:2}$, for $1 \leqq j \leqq p$.

In this instance, it will come from Lie differentiation of a term of the form

$$\xi^{\#}\,\bar\xi_{\eta:\iota_{\eta:k}}\bar\xi^{-1}_{\eta:\iota_{\eta:k}+1}. \tag{10}$$

By induction assumption, (10) has a numerical coefficient that calculates back from our projected coefficient for (6) as

$$-\frac{\iota_{\eta:k}!\,e_{\eta:k}}{(\iota_{\eta:k}-1)!\,e^{\#}}\left(\frac{1}{w}\right)\left(\frac{w!}{\prod_{j=1}^{p}\prod_{k=1}^{r_j} e_{j:k}!\,i_{j:k}!^{e_{j:k}}}\right).$$

Here we take

$$e^{\#} := \begin{cases} e_{\eta:k-1}+1 & \text{if } \iota_{\eta:k-1} = \iota_{\eta:k}-1, \\ 1 & \text{otherwise.} \end{cases}$$

The differentiation process contributes an extra factor of $e^{\#}$. Thus, the partial numerical contribution from this term is

$$\iota_{\eta:k} e_{\eta:k}\left(\frac{1}{w}\right). \tag{11}$$

In this case, the "α" part of the coefficient is carried through unchanged. Moreover, it was already of the required form, since e_η has not been altered by adding and subtracting 1 in (10).

Case 2. It may have an increment in the coefficient of $\bar{\xi}_{j:2}$ for $j = \eta$.

The source of the terms of this type is partial differentiation of the "α" coefficient by $\bar{y}_\eta$. This will give an additional factor of $\bar{\xi}_{\eta:2}$, while not affecting the numerical coefficient. The prior exponent of $\bar{\xi}_{\eta:2}$ will have been $e_{\eta:1} - 1$.

Therefore, the monomial term prior to Lie differentiation was $\xi^{\#} \bar{\xi}_{\eta:2}^{-1}$. By induction assumption, its numerical contribution was

$$-\iota_{\eta:1}!\, e_{\eta:1}\left(\frac{1}{w}\right)\left(\frac{w!}{\prod_{j=1}^{p} \prod_{k=1}^{r_j} e_{j:k}!\, i_{j:k}!^{e_{j:k}}}\right).$$

Its partial numerical contribution is therefore

$$\iota_{\eta:1}\, e_{\eta:1}\left(\frac{1}{w}\right), \tag{12}$$

since $\iota_{j:1}! = 1 = \iota_{j:1}$ when $j = \eta$.

Note that in this case the e_η increments by 1, so that the "α" term adjusts as prescribed by formula (9).

These two cases exhaust the possibilities. The "α" coefficients are as required. And, combining (11) and (12), we get a total contribution to the numerical coefficient of a factor of

$$\left(\sum_{j=1}^{p} \sum_{k=1}^{r_j} \iota_{j:k}\, e_{j:k}\right)\left(\frac{1}{w}\right) = 1,$$

which is also as required. □

The case where not all the observability indices are equal occasions an additional computational process. In this case, we may successively extend the systems, by prolongation, to systems with more indices equal. Thus, if there are three unequal observability indices $l_1 < l_2 < l_3$, we prolong to a system with two observability indices equal to l_2 and one to l_3, of dimension $2l_2 + l_3$, and then to another system with three observability indices, all equal to l_3, of dimension $3l_3$. So doing, we may prolong a given system with arbitrary observability indices to a system with all observability indices equal that is, "equivalent" in the sense that the solutions for $y = y(\bar{y})$ and $\alpha(\cdot)$ can be pulled back trivially to the original, lower dimensional system. This amounts, in effect, to an algorithm for computing the relations between the $\bar{a}$ coefficients and the α injection functions.

The key to this process, is the following lemma. It was originally stated by [14], but the proof there is flawed. The version given here was first proved by Phelps in [17]. The proof is technical and somewhat tedious, and is left for the Appendix.

Note that the lemma is stated in superfluous generality for the purposes of this section. This is so it can also be used to sustain some results of § 3.

LEMMA 2.3. *Suppose an uncontrolled system S, given in observable form coordinates* (ξ, y) (2), *has two distinct multi-indices* l_1, l_2 *of multiplicities* p_1, p_2 *and, further, that S* *may be transformed by change of output coordinates $y = y(\bar{y})$ and change of state coordinates $\xi = \xi(x)$ to observer form coordinates $(x, \bar{y})$. Then S can be prolonged to a system S', also given in observable form* (2), *having multi-indices $\lambda_1 := l_1 + 1$ and $\lambda_2 := l_2$, of the above multiplicities. Furthermore, the transformation $y = y(\bar{y})$ and the function $\alpha(\cdot)$, as given in observer form coordinates, prolong trivially to functions which will take the prolonged system S' over into prolonged observer form.*

We now describe the recursive algorithm for computing polynomial coefficients from injection functions, in the case where the observability indices are not all identical. Lemma 2.3 enables us to prolong a system with two distinct multi-indices $l_1 < l_2$ to a system p_1 dimensions higher by increasing the dimensionality of each of the l_1-dimensional output subsystems to l_2 dimensions, using in the process $l_2 - l_1$ iterations. The algorithm assumes that this has been done, and computes its effect on the functions $\bar{f}_j(\bar{\xi})$ for $p_1 + 1 \leqq j \leqq p_1 + p_2$. If there are more than two distinct multi-indices, we do this process for the lowest two, then for them, taken together, plus the next lowest one, etc.

Arrange the p observability indices into r sets of distinct multi-indices, where $p_1 = k_1$ and $p_i = k_i - k_{i-1}$ for $2 \leqq i \leqq r$. We have Algorithm 2.4.

ALGORITHM 2.4 (Coefficient Prolongation Algorithm). When we have arbitrary observability indices $l_1, \cdots, l_p$, the following steps enable us to compute the coefficients of the polynomials $\bar{f}_j(\bar{\xi})$, for $1 \leqq j \leqq p$:

1 Compute the polynomials $\bar{f}_j(\bar{\xi})$ for 1 or $k_{i-1} + 1 \leqq j \leqq k_i$ and $1 \leqq i \leqq r$, according to the formula of Theorem 2.2, as applied to systems with k_i observability indices equal to l_{k_i}.

2 The first p_1 of the l_j's, those that are equal to l_1, correspond to valid $\bar{f}_j$'s.

3 Suppose that for the first k_i of these l_j's, we have adjusted to get valid $\bar{f}_j$'s, where $i \leqq r$. We adjust the next p_{i+1} of them.

4 Substitute $L_{\bar{f}}^{s-1} \bar{f}_j$ for $\bar{\xi}_{j:l_j+s}$, where $1 \leqq s \leqq l_{k_{i+1}} - l_j$ and $p_i + 1 \leqq j \leqq p_{i+1}$, eliminating the $\bar{\xi}$'s in $\bar{f}_j$, gained by prolongation.

5 Backsubstitute the solutions for the $\bar{f}_j$'s, where $1 \leqq j \leqq p_i$.

6 The coefficients of the monic monomials in $\bar{f}_j$, $k_i + 1 \leqq j \leqq k_{i+1}$ extend our coefficient solutions to cover the variables indexed by $1 \leqq j \leqq k_{i+1}$.

Proof. This is just a restatement of Lemma 2.3 as an algorithm. □

We formulate the Coefficient Prolongation Algorithm as a theorem, and establish that the relations thus given can be backsolved for the α injection functions.

THEOREM 2.5. *Suppose we have arbitrary observability indices $l_1, \cdots, l_p$. Then*

(i) *We may derive the coefficients $\bar{a}_{m:\dots}(\bar{y})$ in terms of the $\alpha(\bar{y})$ injection terms by application of the Coefficient Prolongation Algorithm.*

(ii) *We may solve these equations for the injection terms $\alpha_{m:*}(\bar{y})$ by integrating an expression in $\bar{a}_{k:1}$ and the coefficients $\bar{a}_{k:\dots}(\bar{y})$ of the "linear" monomials $\bar{\xi}_{k:j}$, $2 \leqq j \leqq l_k$, belonging to $\bar{f}_k(\bar{\xi})$, for $1 \leqq k \leqq m$.*

Proof. Outline. Part (ii) requires a somewhat extended argument. We employ the Coefficient Prolongation Algorithm 2.4 and use a double induction on the index l_m and the time derivative level j, $1 \leqq j \leqq l_m$. In each situation we keep track of the sources of the coefficients and their possible forms.

Body of proof. (i) This has already been done in Theorem 2.2 for the case where all observability indices are the same. The Coefficient Prolongation Algorithm enables us to extend this result inductively whenever $l_h < l_{h+1}$ for $1 \leqq h \leqq p - 1$.

(ii) This, too, is immediate from Theorem 2.2 when all observability indices are equal. The coefficient of the "linear" monomial $\bar{\xi}_{m:j}$ is matched to the injection term

$\alpha_{m:l_m-j+1}$, with $\bar{a}_{m:1}$ being matched to $\alpha_{m:l_m}$. The coefficients of the other monomials follow directly by differentiating and/or integrating and/or multiplication by an integer.

Suppose not all observability indices are equal. The $\alpha_{s:*}$'s for $l_1 = \cdots = l_s$ solve as in the one multi-index case.

Now suppose, by induction hypothesis, that we have successfully solved for $\alpha_{k:*}$ by integrating the $\bar{a}_{k:\ldots}$'s for $1 \leqq k \leqq m-1 < p$. If $l_{m-1} = l_m$, we may exploit the symmetry of the construction to integrate for the $\alpha_{m:*}$'s. Thus, we may assume we are in a prolongation-type situation, and develop the induction step for $k = m$.

Let us again inspect the partial differential equation (p.d.e.) expansion in (5), with $j = m$. Consider the expansion at the monomial $\bar{\xi}_{m:l_m-j+1}$, $1 \leqq j \leqq l_m - 1$, of $\bar{f}_m$. The coefficient $\bar{a}_{m:\cdots|l_m-j+1|\cdots}$ with (m) under the bars can be represented as an expression $\bar{e}_1(\bar{y})$. The Coefficient Prolongation Algorithm provides that this expression may involve (a) $\alpha_{m:*}$'s and their Lie derivatives and (b) expressions deriving from the $\bar{f}_k(\bar{\xi})$'s, for $1 \leqq k \leqq m-1$ and their Lie derivatives. By induction assumption, the type (b) expressions are of the form desired.

Now, we establish a second induction on j, where we solve the expressions $\bar{e}_1(\bar{y})$ for $\alpha_{m:j}$ by (at most) a simple integration of a derivative.

For $j = 1$, the coefficient of $\bar{\xi}_{m:l_m}$ is just $-\partial \alpha_{m:j}/\partial \bar{y}_m$, so that the solution is trivial.

For $j < l_m$, the expression $\bar{e}_1(\bar{y})$ is of the form

$$-\frac{\partial \alpha_{m:j}}{\partial \bar{y}_m} + \bar{e}_2(\bar{y}),$$

where $\bar{e}_2(\bar{y})$ is another expression involving only partials of $\alpha_{m:1}, \cdots, \alpha_{m:j-1}$ and expressions of type (b) above. Hence, it too may be solved by a simple integration of a derivative. Finally, $-\alpha_{m:l_m}$ is an isolated term in the expression for $\bar{a}_{m:1}$, so that we may solve for $\alpha_{m:l_m}$ trivially without benefit of integration, given the induction assumptions.

Hence, we may compute the α's as functions of the $\bar{a}$'s, as required. □

A consequence of Theorem 2.5 is that in standard coordinates $\bar{\xi}$, the polynomial $\bar{f}_m(\bar{\xi})$ will not actually achieve degree l_m. Thus, we will come to view the attainment of degree l_m to be "linked" to the transformation $\bar{\xi} = \bar{\xi}(\xi)$, and hence to $\bar{y} = \bar{y}(y)$ (see § 3). We have Corollary 2.6.

COROLLARY 2.6. *In the standard $\bar{\xi}$ coordinates, the polynomials $f_m(\bar{\xi})$, $1 \leqq m \leqq p$, are of degree at most $l_m - 1$.*

Proof. In the p.d.e. expansion (5) and the general expansion for $f_m(\bar{\xi})$, each of the $\alpha_{j:1}$'s, $1 \leqq j \leqq p$, undergoes the *most* Lie differentiations, at $l_m - 1$ times.

Successive Lie differentiations of any α will increase the degree of the resulting expression by (at most) one each time. Even if we have $l_j < l_m$, and we eventually hit a contribution from $f_j(\bar{\xi})$, it will only increase the degree by (at most) one. Hence, the (theoretically possible) coefficients of the monomials of degree l_m, in $f_m(\bar{\xi})$ must vanish, as required. □

The Coefficient Prolongation Algorithm 2.4 provides a process for computing coefficients in the general case. An explicit formula could surely be developed, but it would be quite complicated. The next corollary to Theorem 2.2 treats the simplest case where not all the observability indices are identical, viz., the generic case, with $n \not\equiv 0 \pmod p$.

COROLLARY 2.7. *For the generic case of two different multi-indices of size λ_1 and $\lambda_2 := \lambda_1 + 1$ and multiplicities p_1 and p_2, the coefficient $\bar{a}_{m:\ldots}(\bar{y})$ is given by*

$$\left(\frac{w!}{\prod_{j=1}^{p} \prod_{k=1}^{r_j} e_{j:k}!\, i_{j:k}!^{e_{j:k}}}\right)\left[\sum_{i=1}^{p_1}\left(\frac{\partial^e \alpha_{i:\lambda_1-w}(\bar{y})}{\partial \bar{y}_1^{e_1}\cdots\partial \bar{y}_p^{e_p}}\,\frac{\partial \alpha_{m:l_m-\lambda_1}(\bar{y})}{\partial \bar{y}_i}\right)-\frac{\partial^e \alpha_{m:\lambda_2-w}(\bar{y})}{\partial \bar{y}_1^{e_1}\cdots\partial \bar{y}_p^{e_p}}\right]. \tag{13}$$

In particular, $\bar{a}_{m:\ldots}(\bar{y})$ is given by Theorem 2.2 *for* $1 \leqq m \leqq p_1$ *and for degree* $\geqq \lambda_1$ *when* $p_1 + 1 \leqq m \leqq p$.

Proof. Again, consider the expansion (5). We need to do exactly one prolongation, where we substitute in effect $\bar{f}_1(\bar{\xi})$ for the "quasi variable" $\bar{\xi}_{1:l_2}$ in the prolonged version (which is identical to the expansion (5)).

That is, take

$$\bar{f}_j(\bar{\xi}) = -\sum_{i=1}^{l_j} L_f^{l_j-i} \alpha_{j:i} \tag{14}$$

in the multi-index (abusive) notation. The "quasi coefficient" of the "quasi variable" $\bar{\xi}_{1:l_2}$, gained from prolongation, satisfies the relation

$$\bar{a}_{2:l_2|} = -\frac{\partial \alpha_{2:1}}{\partial \bar{y}_1}. \tag{15}$$

Now, employ the substitution

$$\bar{\xi}_{1:l_2} = \bar{f}_1(\bar{\xi}). \tag{16}$$

The monomial $\bar{\xi}_{1:l_2}$ has degree $l_2 - 1$. By Corollary 2.6, this is the maximum degree possible permitting a term in $\bar{f}_2(\bar{\xi})$ to have a nonzero coefficient. Hence, this is the only monomial we must replace in the prolongation by a substitution.

Now, consider a monomial in $\bar{f}_2$. Its coefficient may receive contributions from (a) the substitution above for the "quasi variable" $\bar{\xi}_{1:l_2}$ and (b) directly from the expression in the right-hand side of formula (14). Thus, evaluating (14) using (15) and (16), we get that the contributions (a) and (b) correspond exactly to the first and second terms (respectively) in (13), as required. □

These results imply that we have a third way of formulating necessary and sufficient conditions for the existence of observer form coordinates for a system. That is, the necessary change of state coordinates $\bar{\xi} = \bar{\xi}(x)$ exists precisely when the relationship between the $\bar{a}$ coefficients and the α injection functions prescribed by Theorem 2.5 can be arranged. In other words, there are certain dependencies which must be obtained in order for the coefficients to be mutually compatible. We formulate *coefficient compatibility* below.

CONDITION 2.8. (Coefficient Compatibility Condition). The coefficients of the polynomials $\bar{f}_m(\bar{\xi})$, for $1 \leqq m \leqq p$, given in standard coordinates, satisfy the injection relations defined by Theorem 2.5. In particular, the coefficients must be compatible with the dependencies given by these relations.

Summing up, we have Theorem 2.9.

THEOREM 2.9. *The observable form, output coordinate change, and polynomial degree conditions, together with the Coefficient Compatibility Condition* 2.8, *are necessary and sufficient for an uncontrolled system to be transformable to nonlinear observer form.*

Proof. This is immediate, since if the first three conditions are satisfied, we can always convert to standard coordinates. □

Let us now examine various examples of the relations that arise from the Coefficient Compatibility Condition.

We express the $\bar{a}$'s in terms of the α's for simplicity's sake. Note that, according to Corollary 2.6, the coefficients of monic monomials of total degree l_m in $\bar{f}_m(\bar{\xi})$ vanish. Hence, they are omitted from these examples.

Example 2.10. Change of state coordinates, single output, dimension four. We have

$$\bar{a}_1 = -\alpha_4, \qquad \bar{a}_2 = -\frac{d\alpha_3}{d\bar{y}},$$

$$\bar{a}_{22} = -\frac{d^2\alpha_2}{d\bar{y}^2}, \qquad \bar{a}_{222} = -\frac{d^3\alpha_1}{d\bar{y}^3},$$

$$\bar{a}_{23} = -3\frac{d^2\alpha_1}{d\bar{y}^2}, \quad \bar{a}_3 = -\frac{d\alpha_2}{d\bar{y}}, \quad \bar{a}_4 = -\frac{d\alpha_1}{d\bar{y}}. \qquad \square$$

We exhibit the three examples where $p = 2$ and $l_2 = 3$.

Example 2.11. Change of state coordinates, $l_1 = l_2 = 3$. Here, the observability indices are equal, corresponding to Theorem 2.2. We have

$$\bar{a}_{1:1} = -\alpha_{1:3}, \qquad \bar{a}_{2:1} = -\alpha_{2:3},$$

$$\bar{a}_{1:22|} = -\frac{\partial^2\alpha_{1:1}}{\partial\bar{y}_1^2}, \qquad \bar{a}_{2:22|} = -\frac{\partial^2\alpha_{2:1}}{\partial\bar{y}_1^2},$$

$$\bar{a}_{1:|2} = -\frac{\partial\alpha_{1:2}}{\partial\bar{y}_2}, \qquad \bar{a}_{2:|2} = -\frac{\partial\alpha_{2:2}}{\partial\bar{y}_2},$$

$$\bar{a}_{1:|22} = -\frac{\partial^2\alpha_{1:1}}{\partial\bar{y}_2^2}, \qquad \bar{a}_{2:|22} = -\frac{\partial^2\alpha_{2:1}}{\partial\bar{y}_2^2},$$

$$\bar{a}_{1:2|} = -\frac{\partial\alpha_{1:2}}{\partial\bar{y}_1}, \qquad \bar{a}_{2:2|} = -\frac{\partial\alpha_{2:2}}{\partial\bar{y}_1},$$

$$\bar{a}_{1:3|} = -\frac{\partial\alpha_{1:1}}{\partial\bar{y}_1}, \qquad \bar{a}_{2:3|} = -\frac{\partial\alpha_{2:1}}{\partial\bar{y}_1},$$

$$\bar{a}_{1:2|2} = -2\frac{\partial^2\alpha_{1:1}}{\partial\bar{y}_1\partial\bar{y}_2}, \qquad \bar{a}_{2:2|2} = -2\frac{\partial^2\alpha_{2:1}}{\partial\bar{y}_1\partial\bar{y}_2},$$

$$\bar{a}_{1:|3} = -\frac{\partial\alpha_{1:1}}{\partial\bar{y}_2}, \qquad \bar{a}_{2:|3} = -\frac{\partial\alpha_{2:1}}{\partial\bar{y}_2}. \qquad \square$$

Example 2.12. Change of state coordinates, $l_1 = 2$, $l_2 = 3$. This is a generic case with two distinct observability indices, corresponding to Corollary 2.7. We have

$$\bar{a}_{1:1} = -\alpha_{1:2}, \qquad \bar{a}_{2:1} = \alpha_{1:2}\frac{\partial\alpha_{2:1}}{\partial\bar{y}_1} - \alpha_{2:3},$$

$$\bar{a}_{1:2|} = -\frac{\partial\alpha_{1:1}}{\partial\bar{y}_1}, \qquad \bar{a}_{2:2|} = \frac{\partial\alpha_{1:1}}{\partial\bar{y}_1}\frac{\partial\alpha_{2:1}}{\partial\bar{y}_1} - \frac{\partial\alpha_{2:2}}{\partial\bar{y}_1},$$

$$\bar{a}_{1:|2} = -\frac{\partial\alpha_{1:2}}{\partial\bar{y}_2}, \qquad \bar{a}_{2:|2} = \frac{\partial\alpha_{1:1}}{\partial\bar{y}_2}\frac{\partial\alpha_{2:1}}{\partial\bar{y}_1} - \frac{\partial\alpha_{2:2}}{\partial\bar{y}_2},$$

$$\bar{a}_{2:22|} = -\frac{\partial^2\alpha_{2:1}}{\partial\bar{y}_1^2}, \qquad \bar{a}_{2:2|2} = -2\frac{\partial^2\alpha_{2:1}}{\partial\bar{y}_1\partial\bar{y}_2},$$

$$\bar{a}_{2:|22} = \frac{\partial^2\alpha_{2:1}}{\partial\bar{y}_2^2}, \qquad \bar{a}_{2:|3} = -\frac{\partial\alpha_{2:1}}{\partial\bar{y}_2}. \qquad \square$$

Example 2.13. Change of state coordinates, $l_1 = 1$, $l_2 = 3$. This is the simplest nongeneric case, and corresponds to our general Theorem 2.5.

$$\bar{a}_{1:1} = -\alpha_{1:1},$$

$$\bar{a}_{2:1} = -\alpha_{1:1}^2 \frac{\partial^2 \alpha_{2:1}}{\partial \bar{y}_1^2} - \alpha_{1:} \frac{\partial \alpha_{1:1}}{\partial \bar{y}_1} \frac{\partial \alpha_{2:1}}{\partial \bar{y}_1} + \alpha_{1:1} \frac{\partial \alpha_{2:2}}{\partial \bar{y}_1} - \alpha_{2:3},$$

$$\bar{a}_{2:|2} = 2\alpha_{1:1} \frac{\partial^2 \alpha_{2:1}}{\partial \bar{y}_1 \partial \bar{y}_2} + \frac{\partial \alpha_{1:1}}{\partial \bar{y}_2} \frac{\partial \alpha_{2:1}}{\partial \bar{y}_1} - \frac{\partial \alpha_{2:2}}{\partial \bar{y}_2},$$

$$\bar{a}_{2:|22} = -\frac{\partial^2 \alpha_{2:1}}{\partial \bar{y}_2^2}, \qquad \bar{a}_{2:|3} = -\frac{\partial \alpha_{2:1}}{\partial \bar{y}_2}. \qquad \square$$

Space problems preclude the exposition of the controlled case (1). In general, we find that the expansion (4) extends naturally to this case. We use this to get results akin to Theorem 2.2, etc. (see [17]).

3. Polynomial coefficients given in observable form coordinates. We found explicit coefficient solutions in the case of standard coordinates. It would be preferable to get explicit solutions in the general observable form case, so that no change of state coordinates will be required. It turns out that this can be readily achieved.

The main adjustment will come from expressions in the parameters of the d.e.'s for change of output coordinates. We will have more degrees of freedom in the choice of polynomial coefficients available to us.

We can convert the results in standard coordinates $(\bar{\xi}, \bar{y})$ to observable form coordinates (ξ, y). For instance, when $p = 1$, we have the change of output coordinates formula

$$\bar{y} = \int_{y^0}^{y} \exp\left(-\frac{1}{l}\int_{y^0}^{\nu} a_{2l}(\eta)\, d\eta\right) d\nu. \tag{17}$$

Repeated Lie differentiations (with backsubstitutions) enable us to compute explicitly the relation $\bar{\xi} = \bar{\xi}(\xi)$.

Using *Macsyma*, this can be done economically for $l < 10$, or so. Let us consider the difficulties in this calculation. We again write J for the change of output coordinates function $dy/d\bar{y}$.

Example 3.1. $\bar{\xi} = \bar{\xi}(\xi)$ calculation, dimension three. We extend the calculation given in Example 2.1 to observable form coordinates ξ. Start with the expansion (7), time differentiate ξ_3, and substitute into (8). With some work, we get

$$a_1 = -J\alpha_3, \qquad a_2 = -J\frac{d\alpha_2}{dy},$$

$$a_{22} = -J\frac{d^2\alpha_1}{dy^2}, \qquad a_{222} = \frac{1}{3}\frac{da_{23}}{dy} - \frac{1}{9}a_{23}^2,$$

$$a_3 = -J\frac{d\alpha_1}{dy}.$$

Also, the x's solved for ξ's come out as

$$x_1 = \int_{y_0}^{y} \exp\left(-\frac{1}{3}\int_{\nu^0}^{\nu} a_{23}\, d\eta\right) d\nu,$$

$$x_2 = J^{-1}\xi_2 - \int_{y^0}^{y} a_3\, d\nu + \alpha_1^0,$$

$$x_3 = J^{-1}\left(\xi_3 - \frac{1}{3}a_{23}\xi_2^2 - a_3\xi_2\right) - \int_{y^0}^{y} a_2\, d\nu + \alpha_2^0. \qquad \square$$

However, it is also possible to find a formula for the a coefficients directly.

The expansion (4) is quite complicated, as expressed in the ξ coordinates. But, once we have worked out the expansion to the $(j-1)$th time derivative, we find that computing from there the jth time derivative is not so complicated. Thus, we have the following theorem.

THEOREM 3.2. *Suppose a single-output system is given in observable form coordinates* (ξ, y), *with index* l. *Then the polynomial* $f_1(\xi)$ *decomposes to*

$$q_l(\xi) - Jp_l(\xi), \tag{18}$$

where $p_l(\xi)$ *is (the negative of) the formal polynomial of degree* $l-1$ *in* ξ, *having the same coefficients as those of* $\bar{f}_1(\bar{\xi})$, *given by Theorem* 2.2, *and* $q_l(\xi)$ *derives from the recursion relations*

$$\begin{aligned} &q_1(\xi) = 0, \\ &q_j(\xi) = L_f q_{j-1}(\xi) + \frac{a_{2l}}{l}(\xi_2(\xi_j - q_{j-1})). \end{aligned} \tag{19}$$

Observe that $q_j(\xi)$ is an abuse of notation, since this polynomial depends formally on l also.

Proof. Outline. The right-hand sides of the expansions in (4) for $f_1(\xi)$ when $j = l$ and for $\xi_{j+1} = L_f\xi_j$ when $j < l$ differ only in the value of the constant l and in the presence, when $j < l$, of the term Jx_{j+1}. We do a simultaneous induction on these two expansions and perform the appropriate backsubstitution. The result matches the recursion relation given by (19).

Body of proof. Note that

$$L_f p_j(\xi) = p_{j+1}(\xi) - \alpha_{j+1}(y) \quad \text{for } j < l. \tag{20}$$

This construction is *formally* identical to the corresponding construction for $\bar{f}$ and $p_j(\bar{\xi})$ under these conditions. First, suppose $j = l = 2$. Direct calculation, using repeated Lie differentiation, gives us

$$\xi_1 = x_1, \qquad J^{-1}\xi_2 = x_2 - \alpha_1(y),$$
$$\xi_2 = J(x_2 - p_1(\xi)),$$

using $p_1(\xi) = \alpha_1(\xi_1) = \alpha_1(y)$. Moreover, using the o.d.e. for J and backsubstituting the previous relation, we continue and get

$$\begin{aligned} f_1(\xi) &= L_f J(x_2 - p_1(\xi)) + J(-\alpha_2(y) - p_2(\xi) + \alpha_2(y)) \\ &= \frac{a_{22}}{2}\xi_2(J(x_2 - p_1(\xi))) - Jp_2(\xi), \\ &= \frac{a_{22}}{2}\xi_2^2 - Jp_2(\xi). \end{aligned}$$

A similar calculation for the case $j = 2$ and $l > 2$ gives us, likewise,

$$\xi_3 = \frac{a_{2l}}{l}\xi_2^2 + J(x_3 - p_2(\xi)).$$

Thus, we evaluate

$$q_2(\xi) = \frac{a_{2l}}{l}\xi_2^2.$$

Since $q_1(\xi)=0$, we see that the relation (19) holds for the case $j=2$. Assume the expansions (4) evaluate to (18) for $j=l$ and simultaneously to

$$q_j(\xi)+J(x_{j+1}-p_l(\xi))$$

for $j<l$, where $q_j(\xi)$ satisfies the recursion relation (19).

We calculate (assuming, first, that $j=l-1$)

$$\begin{aligned}\xi_{j+1}&=q_j(\xi)+J(x_{j+1}-p_j(\xi)),\\ L_f\xi_{j+1}&=L_fq_j(\xi)+L_fJ(x_{j+1}-p_j(\xi))+J(L_fx_{j+1}-L_fp_j(\xi))\\ &=L_fq_j(\xi)+\frac{a_{2l}}{l}\xi_2(J(x_{j+1}-p_j(\xi)))+J(-\alpha_{j+1}(y)-p_{j+1}(\xi)+\alpha_{j+1}(y))\\ &=L_fq_j(\xi)+\frac{a_{2l}}{l}\xi_2(\xi_{j+1}-q_j(\xi))-Jp_{j+1}(\xi).\end{aligned}$$

From this we extract the homogeneous polynomial of degree l and get

$$q_{j+1}(\xi)=L_fq_j(\xi)+\frac{a_{2l}}{l}\xi_2(\xi_{j+1}-q_j(\xi)). \tag{21}$$

Finally, if $j<l-1$, an analogous calculation establishes the relation (21), as required. □

Along the lines of the notation used in § 2 above, we use i_k, e_k, r, e, and w to describe the monomials

$$\xi^{\#}:=\prod_{k=1}^{r}\bar{\xi}_{i_k+1}^{e_k}$$

of $f_1(\xi)$.

In addition, we write $P(m)$ for the partitions of the integer m. We write a partition of $e-1$ by

$$e-1=\sum_{j=1}^{s}c_jn_j,$$

where the n_j's are a set of natural numbers in strictly increasing order. We define $c:=\sum_{j=1}^{s}c_j$ to be the *number of pieces* of the partition. For notational convenience, we also write $d^0a_{2l}(y)/dy^0$ as a synonym for $a_{2l}(y)$.

COROLLARY 3.3. *For the case of a single index l,*

(i) *For monomials of degree less than l, the polynmial coefficient $a_{\dots}(y)$ of $\xi^{\#}$ is given by*

$$-\left(\frac{w!}{\prod_{k=1}^{r}e_k!\,i_k!^{e_k}}\right)\frac{d^e\alpha_{l-w}(y)}{dy^e}\frac{dy}{d\bar{y}}. \tag{22}$$

(ii) *For monomials of degree l, the polynomial coefficient $a_{\dots}(y)$ of $\xi^{\#}$ is given by*

$$-\sum_{\pi\in P(e-1)}\left(\frac{w!(e-1)!}{\prod_{k=1}^{r}e_k!\,i_k!^{e_k}\prod_{j=1}^{s}c_j!n_j!^{c_j}}\right)\left(\frac{-1}{l}\right)^{c}\prod_{j=1}^{s}\left(\frac{d^{n_j-1}a_{2l}(y)}{dy^{n_j-1}}\right)^{c_j}. \tag{23}$$

Proof. Part (i) is immediate from Theorem 3.2.

(ii) We apply the coefficient relation (19). □

Example 3.4. The homogenous polynomial $q_j(\xi)$, for $1\leqq j\leqq 4$, and $l\geqq j$. These are readily generated by direct computation using *Macsyma*, or from formula (23).

We have, writing a_{2l} for $a_{2l}(y)$, etc.,

$$q_1 = 0, \qquad q_2 = \frac{a_{2l}}{l}\xi_2^2,$$

$$q_3 = 3\frac{a_{2l}}{l}\xi_2\xi_3 + \left(\frac{da_{2l}}{dy} - \frac{1}{l^2}\frac{d^2a_{2l}}{dy^2}\right)\xi_2^3,$$

$$q_4 = 4\frac{a_{2l}}{l}\xi_2\xi_4 + 3\frac{a_{2l}}{l}\xi_3^2 + 6\left(\frac{1}{l}\frac{da_{2l}}{dy} - \frac{a_{2l}^2}{l^2}\right)\xi_2^2\xi_3$$

$$+\left(\frac{1}{l}\frac{d^2a_{2l}}{dy^2} - 3\frac{a_{2l}}{l^2}\frac{da_{2l}}{dy} + \frac{a_{2l}^3}{l^3}\right)\xi_2^4. \qquad \square$$

We conjecture that Theorem 3.2 can be readily extended to the case where $p > 1$.

4. Implications and conclusions. We do not have a comprehensive understanding of nonlinear observers. The normal form approach was improved on here and brought closer to effective computability. The primary implication of the coefficient formulas which govern change of state coordinates for nonlinear observers is that these calculations are more susceptible to numeric algorithms than heretofore supposed.

Let us go back and review the bracket vanishing and bracket spanning algorithms. Using the Jacobi identity enables us to reduce the calculation of brackets of elements in $\{\mathrm{ad}_{-f}^{i-1}q_j : 1 \leqq i \leqq l_j\}$ to the calculation of brackets of the form $[\mathrm{ad}_{-f}^{i-1}q_j, q_k]$, for $1 \leqq i \leqq 2l_k - 2$, with indices numbered so that $j \leqq k$. This is a massive saving, but it does not obviate the need for extensive bracket computations in the bracket vanishing case. Similarly, the bracket spanning case (if and when it is extended to the multi-index situation) will likely require bracket calculations to the order of the indices.

One solution would be to write algorithms for bracket calculation that will make these calculations faster. Grossman and Larson [6] have made some progress with this. Another solution to this is to provide the explicit solutions to the change of state coordinates differential equations, as provided here. For instance, the author has written a *Macsyma* program, "Nonlinear Observer," described in [17], which computes the observer *abstractly* and *in the general case* when $p = 1$ and is extendable to the multiple output case. A third approach (see [7], [11]) now being tried involves "approximate normal forms." All these advances indicate that numeric algorithms now seem to be more closely at hand.

Another point worthy of note is that the coefficient compatibility approach implies a corresponding parameter estimation problem [3]. Suppose we have a system which we presume can be put in nonlinear observable form (as is generically the case). Suppose it satisfies the degree condition (this depends on the type of system under consideration). Then the system can be estimated by estimating the $a(y)$ coefficients.

If the system admits nonlinear observer form, then Theorem 2.5, as postulated to extend to the general case for output coordinates, says that the polynomial coefficients for the observable form polynomials are all derivatives of the $\alpha(y)$ injection functions. That is, the injection functions may be found by integration of a certain subset of the $a(y)$ coefficients. Either way, the system is determined by a (small) finite set of parameters.

In the dimension three example (2.1), the system is determined by the parameters $a_1(y)$, $a_2(y)$, and $a_3(y)$, which are the coefficients of the *affine* terms. The incorporation of the change of outputs coordinates (Thm. 3.2) gives us one additional parameter, the $a_{23}(y)$ coefficient that determines the o.d.e. governing this coordinate change. (Note how the parameters in nonlinear observer form extend those of the linear case!)

The number of parameters is affected by increasing the number of output coordinates p, since then the a's become functions of more variables.

In short, there is room to develop a theory of parameter estimation and state identification for systems which may be put into observer normal form.

It is important to do research in and to develop a full account of nonlinear observers. We have not attempted to go that far, but hopefully this contribution will provide some new impetus and direction for pursuing this goal.

Appendix A. Proof of the prolongation lemma. We present the proof of Lemma 2.3. First, we require Proposition A1.

PROPOSITION A1. *For a system with arbitrary indices, it is necessary that* $\partial y_j/\partial \bar{y}_i =: J_i^j = 0$ *when* $l_j > l_i$.

Proof. The proof is immediate. See [14] or [17] for details. □

Proof of Lemma 2.3. *Outline.* Essentially, we add a new multidimension, by way of the coordinate $\check{x}_{1:\lambda_1}$. The solution to S is to be prolonged to the hyperplane

$$\check{x}_{1:\lambda_1} = 0. \tag{24}$$

We set up the prolongation relations and verify that S' may be attained.

Body of proof. We exhibit the prolonged system S' in observable and observer forms. By abuse of notation, we are using y_j, $j = 1, 2$, to represent p_j-vectors. We do likewise for the l_j's, $\xi_{j:*}$'s, $x_{j:*}$'s, $\alpha_{j:*}$'s, f_j's, the prolonged coordinates $(\check{\xi}, \check{\bar{y}})$, etc. We have

$$\begin{aligned} \check{y}_j &= \check{\xi}_{j:1}, & \check{\bar{y}}_j &= \check{x}_{j:1}, \\ \dot{\check{\xi}}_{j:k} &= \check{\xi}_{j:k+1}, & \dot{\check{x}}_{j:k} &= \check{x}_{j:k+1} - \alpha_{j:k}(\check{y}), \\ \dot{\check{\xi}}_{j:\lambda_j} &= \check{f}_j(\check{\xi}), & \dot{\check{x}}_{j:\lambda_j} &= -\alpha_{j:\lambda_j}(\check{y}), \end{aligned} \tag{25}$$

with $j = 1, 2$ and $1 \leqq k \leqq \lambda_j - 1$.

Here we are taking

$$\begin{aligned} &\check{\xi}_{j:k} := \xi_{j:k} \quad \text{when } j = 1, 2 \text{ and } 1 \leqq k \leqq \lambda_1 - 1, \\ &\alpha_{1:\lambda_1} := 0, \end{aligned} \tag{26}$$

so that

$$\dot{\check{x}}_{1:\lambda_1} = 0 \tag{27}$$

holds.

To set up the remainder of the prolongation relations, it is advisable to look at the defining p.d.e.'s for S and see how they must be "adjusted" for the prolongation to S'.

Thus, we continue the theme expressed in formula (4), extracting p.d.e.'s relating observable form (2) to observer form (3). We now exhibit that expansion for the two multi-index case under consideration. We arrange the dimensionalities of the "α," "f," "ξ," "x" and "J" notation appropriately and get

$$\begin{aligned} &\xi_{j:1} = y_j(\bar{y}) = y_j(x_1), \\ &\quad \xi_{j:2} = \dot{\xi}_{j:1} = J_j^1(x_{1:2} - \alpha_{1:1}) + J_j^2(x_{2:2} - \alpha_{2:1}), \\ &\quad \xi_{j:3} = \dot{\xi}_{j:2} = \sum_{k=1}^{2} J_j^k\left(x_{k:3} - \alpha_{k:2} - \sum_{r=1}^{2} \frac{\partial \alpha_{k:1}}{\partial \bar{y}_r}(x_{r:2} - \alpha_{r:1})\right) + \cdots, \\ &\qquad \vdots \\ &\xi_{j:l_1} = \dot{\xi}_{j:l_1-1} = \sum_{k=1}^{2} J_j^k\left(x_{k:l_1} - \alpha_{k:l_1-1} - \sum_{r=1}^{2} \frac{\partial \alpha_{k:l_1-2}}{\partial \bar{y}_r}(x_{r:2} - \alpha_{r:1}) + \cdots\right) + \cdots, \end{aligned} \tag{28}$$

with additional terms depending on j.

The completion of (26) for $j=1$ goes as follows.

CLAIM. We can prolong f_1 by the relations

$$\check{\xi}_{1:\lambda_1} := J_1^1 \check{x}_{1:\lambda_1} + f_1(\check{\xi}), \tag{29}$$

$$\check{f}_1(\check{\xi}) := \frac{\partial J_1^1}{\partial y_1} \check{\xi}_{1:2} (J_1^1)^{-1} (\check{\xi}_{1:\lambda_1} - f_1) + \sum_{j=1}^{2} \sum_{k=1}^{l_j} f_{1;j:k} \dot{\check{\xi}}_{j:k}. \tag{30}$$

We may match the continuation of (28) to its prolonged version. We have

$$f_1(\xi) = \dot{\xi}_{1:l_1} = J_1^1 \left(-\alpha_{1:l_1} - \sum_{r=1}^{2} \frac{\partial \alpha_{1:l_1-1}}{\partial \bar{y}_r} (x_{r:2} - \alpha_{r:1}) + \cdots \right) + \cdots, \tag{31}$$

and

$$\begin{aligned} \check{\xi}_{1:\lambda_1} &= \dot{\check{\xi}}_{1:\lambda_1-1} \\ &= J_1^1 \left(\check{x}_{1:\lambda_1} - \alpha_{1:\lambda_1-1} - \sum_{r=1}^{2} \frac{\partial \alpha_{1:\lambda_1-2}}{\partial \bar{y}_r} (\check{x}_{r:2} - \alpha_{r:1} + \cdots \right) + \cdots . \end{aligned} \tag{32}$$

Recall that Proposition $A1$ establishes that $J_1^2 = 0$.

Now, assuming the partial prolongation (26), we may compare (31) (using $\check{\xi}$ and $\check{x}$ for ξ and x) with the corresponding equation (32). We get

$$\check{\xi}_{1:\lambda_1} = J_1^1 \check{x}_{1:\lambda_1} + f_1(\check{\xi}),$$

whence (29).

For our prolongation to be possible, we must have $\dot{\check{\xi}}_{1:\lambda_1} = \check{f}_1(\check{\xi})$. Differentiating (29), and recalling that $\dot{\check{x}}_{1:\lambda_1} = 0$ (27) and $J_1^2 = 0$, we get

$$\dot{\check{\xi}}_{1:\lambda_1} = \frac{\partial J_1^1}{\partial y_1} \xi_{1:2} \check{x}_{1:\lambda_1} + \sum_{j=1}^{2} \sum_{k=1}^{l_j} f_{1;j:k} \dot{\check{\xi}}_{j:k}. \tag{33}$$

Finally, we take $\check{f}_1(\check{\xi}) := \dot{\check{\xi}}_{1:\lambda_1}$, and backsolve in (33) for $\check{f}_1(\check{\xi})$. Since J_1^1 does not vanish locally, we may do this, and the claim follows.

The completion of (26) for $j=2$ is somewhat more intricate. This is due to "twistings" induced in the p.d.e.'s for the higher time derivatives of y_2 by the (formal) introduction of $\check{\xi}_{1:\lambda_1}$.

Extending (28) for the $j=2$ case, we have

$$\begin{aligned} \xi_{2:l_1+1} &= \dot{\xi}_{2:l_1} \\ &= J_2^2 x_{2:l_1+1} + \sum_{k=1}^{2} J_2^k \left(-\alpha_{k:l_1} - \sum_{r=1}^{2} \frac{\partial \alpha_{k:l_1-1}}{\partial \bar{y}_r} (x_{r:2} - \alpha_{r:1}) + \cdots \right) + \cdots, \\ \xi_{2:l_1+2} &= \dot{\xi}_{2:l_1+1} \\ &= J_2^2 (x_{2:l_1+2} + \alpha_{2:l_1+1}) + \left(\sum_{r=1}^{2} \frac{\partial J_2^2}{\partial \bar{y}_r} (x_{r:2} - \alpha_{r:1}) \right) x_{2:l_1+1} \\ &\quad + \sum_{k=1}^{2} J_2^k \left(- \sum_{r=1}^{2} \frac{\partial \alpha_{k:l_1}}{\partial \bar{y}_r} (x_{r:2} - \alpha_{r:1}) + \cdots \right) + \cdots, \\ &\ \ \vdots \\ \xi_{2:l_2} &= \dot{\xi}_{2:l_2-1} = J_2^2 (x_{2:l_2} - \alpha_{2:l_2-1}) + \cdots \\ &\quad + \sum_{k=1}^{2} J_2^k \left(- \sum_{r=1}^{2} \frac{\partial \alpha_{k:l_2-2}}{\partial \bar{y}_r} (x_{r:2} - \alpha_{r:1}) + \cdots \right) + \cdots, \\ f_2(\xi) &= \dot{\xi}_{2:l_2} \\ &= -J_2^2 \alpha_{2:l_2} + \cdots + \sum_{k=1}^{2} J_2^k \left(- \sum_{r=1}^{2} \frac{\partial \alpha_{k:l_2-1}}{\partial \bar{y}_r} (x_{r:2} - \alpha_{r:1}) + \cdots \right) + \cdots . \end{aligned} \tag{34}$$

What occurs when we introduce the new variable $\check{x}_{1;\lambda}$ as in (25)? We redo (34), this time using "checked" coordinates $\check{\xi}$ and $\check{x}$. By abuse of notation, we write "$\xi_{2;j}$," $l_1+1\leqq j\leqq l_2$, to denote the expression given by the corresponding right-hand side in (34), with $\check{x}$'s replacing x's.

We may now take

$$\begin{aligned}\check{\xi}_{2:\lambda_1} &:= s_0\check{x}_{1:\lambda_1}+\xi_{2:l_1+1},\\ \check{\xi}_{2:\lambda_1+1} &:= (L_{\check{A}\check{x}-\alpha}s_0+s_1)\check{x}_{1:\lambda_1}+\xi_{2:l_1+2},\\ &\ \vdots\\ \check{\xi}_{2:\lambda_2} &:= \left(\sum_{\sigma=0}^{\lambda_2-\lambda_1} L_{\check{A}\check{x}-\alpha}^{\lambda_2-\lambda_1-\sigma}s_\sigma\right)\check{x}_{1:\lambda_1}+\xi_{2:l_2},\\ \check{f}_2(\check{\xi}) &:= \left(\sum_{\sigma=0}^{\lambda_2-\lambda_1+1} L_{\check{A}\check{x}-\alpha}^{\lambda_2-\lambda_1+1-\sigma}s_\sigma\right)\check{x}_{1:\lambda_1}+f_2(\check{\xi}).\end{aligned}\tag{35}$$

Here, $\check{A}\check{x}-\alpha$ is $\check{f}$ in the $\check{x}$ coordinates, with $\check{A}$ being in Brunovsky canonical form. α, by abuse of notation, represents the prolonged $(n+1)$-vector

$$\begin{bmatrix}\alpha_{1:1}\\ \vdots\\ \alpha_{1:\lambda_1-1}\\ 0\\ \alpha_{2:1}\\ \vdots\\ \alpha_{2:\lambda_2}\end{bmatrix}.$$

The s_σ's, $0\leqq\sigma\leqq\lambda_2-\lambda_1+1$, are polynomials in $\check{x}$ with functions in $\bar{y}$ for coefficients. (Note that the coefficients of $\check{x}_{1:\lambda_1}$ will be of degree $\geqq 1$ if $\sigma\geqq 2\lambda_1-1$.)

S' is now defined by formulas (26), (27), (29), (30), and (35).

Finally, consider the hyperplane (24) given by $\check{x}_{1:\lambda_1}=0$. Restricted to this hyperplane, the equations for S and S' are identical, with the *same* α's and transformation $y=y(\bar{y})$. Thus, the solutions are the same, and S' prolongs S, as required for the lemma. □

REFERENCES

[1] D. BESTLE AND M. ZEITZ, *Canonical form observer design for non-linear time-variable systems*, Internat. J. Control, 38 (1983), pp. 419-431.

[2] P. BRUNOVSKÝ, *A classification of linear controllable systems*, Kybernetika, 6 (1970), pp. 173-188.

[3] P. EYKHOFF, *System Identification*, John Wiley, London, 1974.

[4] H. FRITZ AND H. KELLER, *Design of nonlinear observers by a two-step transformation*, in Algebraic and Geometric Methods in Nonlinear Control Theory, M. Fliess and M. Hazewinkel, eds., D. Reidel, Dordrecht, the Netherlands, 1986.

[5] R. GROSSMAN AND R. G. LARSON, *Hopf algebraic structures of families of trees*, Tech. Report PAM-377, Center for Pure and Applied Mathematics, University of California, Berkeley, CA, May 1987.

[6] ———, *Labeled trees and the algebra of differential operators*, Tech. Report PAM-368, Center for Pure and Applied Mathematics, University of California, Berkeley, CA, April 1987.

[7] S. KARAHAN, *Higher order linear approximation to nonlinear systems*, Ph.D. thesis, University of California, Davis, CA, 1989.

[8] H. KELLER, *Nonlinear observer design by transformation into a generalized observer canonical form*, University of Karlsruhe, Karlsruhe, Germany, preprint.

[9] H. KELLER, *Nonlinear observer design via two canonical forms*, University of Karlsruhe, Karlsruhe, Germany, preprint.

[10] A. J. KRENER, *The intrinsic geometry of dynamic observations*, in Algebraic and Geometric Methods in Nonlinear Control Theory, M. Fliess and M. Hazewinkel, eds., D. Reidel, Dordrecht, the Netherlands, 1986.

[11] ———, *Poincaré's linearization method applied to the design of nonlinear compensators*, in preparation.

[12] ———, *Normal forms for linear and nonlinear systems*, in Differential Geometry, the Interface between Pure and Applied Mathematics, M. Luksik, C. Martin, and W. Shadwick, eds., American Mathematical Society, Providence, RI, 1986.

[13] A. J. KRENER AND A. ISIDORI, *Linearization by output injection and nonlinear observers*, Systems Control Lett., 3 (1983), pp. 47-52.

[14] A. J. KRENER AND W. RESPONDEK, *Nonlinear observers with linearizable error dynamics*, SIAM J. Control Optim., 23 (1985), pp. 197-216.

[15] 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.

[16] A. S. MORSE, *Structural invariants of linear multivariable systems*, SIAM J. Control Optim., 11 (1973), pp. 446-465.

[17] A. R. PHELPS, *A simplification of nonlinear observer theory*, Ph.D. thesis, University of California, Berkeley, CA, 1987.

[18] M. ZEITZ, *Canonical forms for nonlinear systems*, in Proc. Conference on the Geometric Theory of Nonlinear Control Systems, Wrocław Technical University Press, Wrocław, Poland, 1985.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 3, pp. 535-561, May 1991

ASYMPTOTIC BEHAVIOR OF STOCHASTIC SYSTEMS POSSESSING MARKOVIAN REALIZATIONS*

S. P. MEYN† AND P. E. CAINES‡

Abstract. Markovian stochastic systems of the form $\Phi_{k+1} = F(\Phi_k, w_{k+1})$ are considered, where $\mathbf{w}$ is an independent and identically distributed process and $\mathbf{\Phi}$ is a Markov chain, both evolving in Euclidean space. A condition called weak stochastic controllability (w.s.c.) is introduced, which is a generalization of the concept of controllability in linear system theory. Two different formulations of stability (boundedness in probability and boundedness in probability on average) are presented, and it is shown that these conditions are equivalent under the w.s.c. condition. These results are related to the ergodic behavior of $\mathbf{\Phi}$ by a set of results that includes the following: if $\mathbf{\Phi}$ is bounded in probability and w.s.c., then (i) sample path averages of functions of the state process converge for every initial condition, and (ii) if the stability is uniform then state process probabilities converge to a periodic orbit of probabilities. Finally, if the origin of an undriven system is globally attracting, the linearized system is controllable, and some technical conditions hold, then it is shown that $\mathbf{\Phi}$ is w.s.c.

Key words. nonlinear systems, accessibility, stochastic systems, Markov chains.

AMS(MOS) subject classifications. 60J05, 93B05, 93E03, 93E15

1. Introduction. The ergodic theory of Markov chains is a natural tool to apply in the analysis of nonlinear stochastic systems under feedback control. Such techniques can be found in the pioneering studies of stochastic stability [Kushner, 1967], [Kushner, 1972], and [Wonham, 1966], and there has recently been a resurgence of interest in their application in such work as [Kushner and Schwartz, 1984], [Kumar, 1985], [Aloneftis, 1987], [Meyn and Caines, 1987], [Meyn and Guo, 1990], Chapter 11 of [Caines, 1988], [Fernandez-Gaucherand, Arapostathis, and Marcus, 1988], and [Arapostathis and Marcus, 1990]. This is partly because the complexity of the feedback systems arising in stochastic control (in particular in stochastic adaptive control) often makes exact analysis impossible, so we seek techniques that will give overall qualitative measures of behavior.

Ergodic theory provides one such approach for stochastic systems once a Markovian state process has been constructed for the controlled output process. In the case where an appropriate form of stability holds together with other technical conditions, we may deduce (i) the existence of an invariant measure π for the process, and (ii) the convergence almost surely of the sample averages of a function of the state process (and its expectation) to its conditional expectation with respect to the sub σ-algebra of invariant sets.

In the ergodic theory of Markov chains' most general setting, it is not possible to establish the desired ergodic properties unless the initial condition of the process lies in a set of full measure with respect to the invariant probability π. A central issue in this paper is generalizing such results to the case of arbitrary initial conditions.

* Received by the editors May 17, 1988; accepted for publication (in revised form) March 8, 1990. This work was conducted at the Department of Electrical Engineering, McGill University; the Department of Systems Engineering, Australian National University; and the Coordinated Science Laboratory, University of Illinois. This research was supported in part by Natural Sciences and Engineering Research Council of Canada and National Science Foundation grant ECS8910088.

† Department of Electrical Engineering and the Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, Illinois 61801.

‡ Department of Electrical Engineering and the Canadian Institute for Advanced Research, McGill University, Montréal, Quêbec, Canada H3A 2A7.

An example. To motivate the discussion and definitions that follow, consider the Gaussian Markov process $\mathbf{\Phi}$ generated by the recursion

$$\Phi_{k+1} = A\Phi_k + Bw_{k+1}, \tag{1}$$

where $\mathbf{\Phi}$ and $\mathbf{w}$ evolve on $\mathbb{R}^n$ and $\mathbb{R}^p$, respectively; A and B are, respectively, $n \times n$ and $n \times p$ matrices; $\mathbf{w} = \{w_k : k \geqq 1\}$ is an independent and identically distributed 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 $\mathbb{C}$. Then many of the asymptotic properties of (1) are determined by its unique invariant probability π. The probability π is the Gaussian distribution $N(0, F)$, where F is the unique solution to the Lyapunov equation

$$F = AFA^\top + BB^\top.$$

If the pair (A, B) is controllable then an analysis of the asymptotic properties of $\mathbf{\Phi}$ is straightforward. If f is any positive Borel function on $\mathbb{R}^n$, then by the Strong Law of Large Numbers for Markov chains [Doob, 1953], for almost every initial condition $\Phi_0 = x \in \mathbb{R}^n[\pi]$,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(\Phi_k) = \int f\, d\pi \quad \text{a.s. } [P_x], \tag{2}$$

and by a simple computation,

$$\lim_{k\to\infty} E_x[f(\Phi_k)] = \int f\, d\pi. \tag{3}$$

By conditioning at time n, where $n = \dim \mathbf{\Phi}$, (2) may be generalized to arbitrary initial conditions by using the fact that $P^n(x, \cdot)$ is absolutely continuous with respect to π, which is implied by the controllability assumption. Hence if (1) describes a stochastic system operating under feedback, and f is a loss function on the state process $\mathbf{\Phi}$, then by (2), (3) the infinite horizon performance is determined by the invariant probability π.

On the other hand, if (A, B) is not controllable then (2), (3) does not hold for such a general class of functions in general. Because the covariance matrix F is not full rank in this case, the invariant probability π is supported on the controllability subspace $L \subset \mathbb{R}^n$ whose dimension is strictly less than n. Hence (2), (3) may not hold unless f is continuous on L. (Take, for example, f to be the indicator function of the set L.) To establish (2), (3) even for continuous functions requires extensive exploitation of the linear structure of (1). Motivated by such considerations our objective in this paper is to generalize the notion of controllability to analyse nonlinear stochastic systems operating under feedback.

Methodology and previous results. The principal tools applied (such as Harris recurrence) come from the theory of irreducible Markov chains; see for example [Nummelin, 1984]. It turns out that some form of stochastic controllability condition is precisely what is needed to obtain irreducibility for the state process $\mathbf{\Phi}$, and then a crude stability hypothesis implies the ergodic theorems of interest. In general, when no controllability hypothesis is satisfied, it is natural to search for a restricted class of functions (say, continuous bounded functions) for which (2), (3) hold. However, with the exception of the results for the class of *regular* Markov chains as presented in [Feller, 1971] that are extremely difficult to apply since a general verifiable criterion for regularity is not available, and the theory presented in [Foguel, 1973] that concentrates on the structure of the state space rather than stability and ergodic theory, there is no alternative framework available that is suitable for the analysis of stochastic control systems.

The present paper may be compared to [Kliemann, 1987] (see also [Ichihara and Kunita, 1974]), where diffusions possessing hypoelliptic generators are considered. However, the so-called crucial use of continuity of the samples paths of the processes under consideration, and use of the resolvant operator of the process, make these results meaningless in a discrete time setting. The techniques of the present paper extend naturally to the nonsmooth case, and may also be used to considerably strengthen the results now available in the continuous time Markov process literature.

Generalizations of some of the results of this paper to the case of discontinuous dynamics, and to continuous time processes with discontinuous sample paths (including Hunt processes) has begun in [Meyn and Tweedie, 1990a,b].

Overview of results. The paper is organized as follows. In § 2 we begin with definitions of weak and local stochastic controllability, discuss further formulations of controllability for a freely evolving system, and present some stability criteria for Markovian systems including boundedness in probability, and an averaged formulation of this condition. The main results of the paper are then presented: if $\mathbf{\Phi}$ is bounded in probability and w.s.c., then (i) sample path averages of functions of the state process converge for every initial condition, and (ii) if the stability is uniform, then state process probabilities converge to a periodic orbit of probabilities. Finally, if the origin of an undriven system is globally attracting, the linearized system is controllable, and some technical conditions hold, then it is shown that $\mathbf{\Phi}$ is w.s.c.

In § 3 some general results concerning the topological structure of the state space are derived and then applied to establish limit theorems for weakly stochastically controllable systems and hence the proofs of the main results.

2. Markovian systems. In this paper we consider input-state stochastic systems possessing Markovian realizations (which we call Markovian systems) of the form

$$\Phi_{k+1}=F(\Phi_k, w_{k+1}), \qquad k\in\mathbb{Z}_+, \tag{4}$$

where for all k, $\Phi_k\in\mathbf{X}=$ an open subset of $\mathbb{R}^n$, $w_k\in\mathbb{R}^p$, and $F:\mathbf{X}\times\mathbb{R}^p\to\mathbf{X}$ is smooth (C^∞).

We assume that the initial condition Φ_0 and the distrubance process $\mathbf{w}$ satisfy

A1. $(\Phi_0, \mathbf{w})$ are random variables on the probability space $(\Omega, \mathcal{F}, P_{\Phi_0})$;

A2. Φ_0 is independent of $\mathbf{w}$;

A3. $\mathbf{w}$ is an independent and identically distributed process;

A4. The distribution μ_w of w_k, $k\in\mathbb{Z}_+$, possesses a density that is lower semicontinuous.

A function $f:\mathbb{R}^p\to\mathbb{R}$ is *lower semicontinuous* if the set $\{x\in\mathbb{R}^p: f(x)>t\}$ is open for each $t\in\mathbb{R}$. The Vitali-Carathéodory Theorem implies that any $f\in L^1(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), dx)$ may be approximated in L^1 by lower semicontinuous functions, and hence any distribution that possesses a density may be approximated in total variation norm by distributions satisfying condition A4.

Condition A4 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.

Markovian systems will be obtained from input-state-output systems by the choice of time invariant feedback control laws. To obtain the ergodic properties of interest for $\mathbf{\Phi}$ it will, of course, be necessary to verify that each particular feedback law generates a system satisfying the appropriate hypotheses.

The crucial property of $\mathbf{\Phi}$ that permits us to obtain these ergodic properties is that it possesses stationary Markovian transition probabilities P^k, $k \in \mathbb{Z}_+$, satisfying the defining property

$$E[f(\Phi_{n+k}) \mid \Phi_n] = \int_{\mathbf{X}} P^k(\Phi_n, dy) f(y) \quad \text{a.s. } [P_{\Phi_0}]$$

for all bounded Borel measurable functions f on $\mathbf{X}$. Because F is a continuous function of its arguments, the Markov chain $\mathbf{\Phi}$ has the Feller property.

2.1. Controllability. The *extended transition map* $S_x^k : \mathbb{R}^{kp} \to \mathbf{X}$ of the Markovian system (4) is defined inductively for $k \in \mathbb{Z}_+$, $x \in \mathbf{X}$, and $\mathbf{z} = (z_1, \cdots, z_k)^\top \in \mathbb{R}^{pk}$ by

$$S_x^k(\mathbf{z}) = F(S_x^{k-1}(z_1, \cdots, z_{k-1}), z_k), \qquad k \geqq 1,$$

$$S_x^0 = x.$$

The extended transition map is so named because for all $k \in \mathbb{Z}_+$,

$$\Phi_k = S_x^k(w_1, \cdots, w_k) \quad \text{when } \Phi_0 = x.$$

Stochastic controllability. Here we introduce two useful formulations of stochastic controllability. Given two measures ν and μ on $\mathcal{B}(\mathbf{X})$ we say that ν is *absolutely continuous* with respect to μ (denoted $\nu < \mu$) if $\nu\{A\} = 0$ whenever $\mu\{A\} = 0$. The measures ν and μ are called *equivalent* (denoted $\nu \approx \mu$) if $\nu < \mu$ and $\mu < \nu$. Throughout this paper 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. (i) The Markovian system (4) is called *locally stochastically controllable* (l.s.c.) 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 $P^T(x, \cdot) \approx \mathbf{1}_{\mathcal{O}_x} \mu^{\text{Leb}}$.

(ii) The Markovian system (4) is called *weakly stochastically controllable* (w.s.c.) 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 $P^T(x, \cdot) > \mathbf{1}_{\mathcal{O}_x} \mu^{\text{Leb}}$.

If (4) is l.s.c. then the probability $P^T(x, \cdot)$ possesses a density that is strictly positive on an open set $\mathcal{O}_x$ and zero elsewhere. Similarly, if (4) is w.s.c. then the Radon-Nikodym derivative of the probability $P^T(x, \cdot)$ (with respect to Lebesgue measure) is strictly positive on $\mathcal{O}_x$.

One consequence of these definitions may be roughly described as follows: If (4) is weakly or locally stochastically controllable, and if starting at a point $y \in \mathbf{X}$ it is possible to reach a point $z \in \mathcal{O}_y$ at time T, then at time T all points in some neighborhood of z are reachable from y. In fact, under assumptions A1-A4 and the smoothness condition made on F, it is easily verified that with the disturbance sequence $\mathbf{w}$ viewed as an input, the forward accessibility condition of [Jakubczyk and Sontag, 1989] is equivalent to weak stochastic controllability. The terminology may also be motivated by the fact that if $F : \mathbf{X} \times \mathbb{R}^p \to \mathbf{X}$ is linear then the notions of local stochastic controllability, weak stochastic controllability, and controllability in the usual sense are equivalent.

For $y \in \mathbf{X}$ and a sequence $\{z_k : z_k \in \mathbb{R}^p, k \in \mathbb{Z}_+\}$ let $\{A_k, B_k : k \in \mathbb{Z}_+\}$ denote the matrices

$$A_k = A_k(y, z_1, \cdots, z_{k+1}) \triangleq \left[\frac{\partial F}{\partial x}\right]_{(S_y^k, z_{k+1})}$$

$$B_k = B_k(y, z_1, \cdots, z_{k+1}) \triangleq \left[\frac{\partial F}{\partial z}\right]_{(S_y^k, z_{k+1})},$$

and let $C_y^k = C_y^k(z_1, \cdots, z_k)$ denote the *generalized controllability matrix* (along the sequence z_1^k)

$$C_y^k \triangleq [A_{k-1} \cdots A_1 B_0 | A_{k-1} \cdots A_2 B_1 | \cdots | A_{k-1} B_{k-2} | B_{k-1}]. \tag{5}$$

We remark that if F is of the form

$$F(y, z) = Ay + Bz,$$

then the generalized controllability matrix becomes the familiar controllability matrix

$$[A^{T-1}B | A^{T-2}B | \cdots | AB | B].$$

Note that all quantities in the matrix (5) are deterministic.

Here we give necessary and sufficient conditions for local and weak stochastic controllability in terms of the generalized controllability matrix defined above. Alternative conditions for weak stochastic controllability involving the dimension of a certain Lie algebra and substantially stronger conditions on the function F may be found in [Jakubczyk and Sontag, 1989].

Let $\mathcal{O}_w^T$ denote the T-fold Cartesian product of $\mathcal{O}_w$ (recall that $\mathcal{O}_w$ is the open set that supports μ_w, defined at the beginning of this section).

THEOREM 2.1. *Suppose that* $\mathbf{\Phi}$ *is of the form* (4) *and that conditions* A1–A4 *hold. Then*

(i) *The Markovian system* (4) *is l.s.c. if and only if for all initial conditions* $x \in \mathbf{X}$ *there exists* $T \geqq 1$ *such that*

$$\text{rank}\, C_x^T(\lambda) = n \quad \textit{for all } \lambda \in \mathcal{O}_w^T \backslash Z_x, \tag{6}$$

where $Z_x \cap \mathcal{O}_w^T$ *has zero Lebesgue measure.*

(ii) *The Markovian system* (4) *is w.s.c. if and only if for all initial conditions* $x \in \mathbf{X}$, *there exists* $T \geqq 1$ *and* $\lambda \in \mathcal{O}_w^T$ *such that*

$$\text{rank}\, C_x^T(\lambda) = n. \tag{7}$$

(iii) *If* rank $C_x^T(\lambda) = n$ *for some* $\lambda \in \mathcal{O}_w^T$, *then there exists* $c > 0$, *and open sets* $\mathcal{U}_x^\lambda$, $\mathcal{V}_x^\lambda$ *containing* x *and* $S_x^T(\lambda)$, *respectively, such that*

$$P^T(y, A) \geqq c\mu^{\text{Leb}}\{A \cap \mathcal{V}_x^\lambda\} \tag{8}$$

for all $A \in \mathcal{B}(\mathbf{X})$ *and* $y \in \mathcal{U}_x^\lambda$.

The proof of Theorem 2.1 is given in § 3. We remark that (6) is equivalent to the condition that the random matrix

$$C_x^T(w_1 \cdots w_T)$$

is full rank almost surely, and (since C_x^T is smooth) (7) is equivalent to the condition that this matrix is full rank with positive probability.

Equation (8) may be written in the symbolic form

$$P^T(\cdot, \cdot) \geqq c\mathbf{1}_{\mathcal{U}_x^\lambda}(\cdot)(\mathbf{1}_{\mathcal{V}_x^\lambda}\mu^{\text{Leb}})\{\cdot\}.$$

A set $\mathcal{U}$ and measure φ satisfying

$$P^T(\cdot, \cdot) \geqq c\mathbf{1}_{\mathcal{U}}(\cdot)\varphi\{\cdot\} \tag{9}$$

for some $c > 0$ are called, respectively, a small set and a small measure [Nummelin, 1984]. It is the existence of open small sets together with the Feller property that allows points of the state space to be connected together. Once it can be shown that the state process $\mathbf{\Phi}$ enters some small set infinitely often with positive probability, it follows that every trajectory of probabilities eventually dominates the small measure φ and Harris recurrence (or a generalization as in Theorems 2.2 and 2.3) follows easily. Most of the stability conditions introduced in this paper will be used to show that an open small set is attracting in this sense.

From the fact that the state space may be covered by open small sets we may deduce that a certain stochastic kernel possesses continuous components, allowing an application of the results of [Tuominen and Tweedie, 1979]. However these results are not needed in the present paper.

By Theorem 2.1 it follows that, in many problems encountered in signal processing and adaptive control problems, the notions of local and weak stochastic controllability are equivalent.

COROLLARY 2.1. *If conditions* A1–A4 *hold and the function F defined in* (4) *is a real-analytic function of its arguments, then* $\mathbf{\Phi}$ *is l.s.c. if and only if it is w.s.c.*

Controllability to a fixed state. Given a Markovian system of the form (4) and a point $w^\star \in \mathbb{R}^p$ we will call the deterministic system

$$d_{k+1} = F(d_k, w^\star), \qquad k \in \mathbb{Z}_+ \tag{10}$$

with initial condition $d_0 \in \mathbf{X}$ the *freely evolving system.*

A5. There exist $w^\star \in \mathbb{R}^p$ and $d^\star \in \mathbf{X}$ such that

$$d^\star \in \overline{\left\{ \bigcup_{k=0}^{\infty} S_x^k(w^\star, \cdots, w^\star) \right\}} \tag{11}$$

for every initial condition $x \in \mathbf{X}$.

Observe that if condition A5 is satisfied then in particular, (11) holds with initial condition $d_N = S_x^N(w^\star, \cdots, w^\star)$. Hence for all $x \in \mathbf{X}$ and $N \in \mathbb{Z}_+$,

$$d^\star \in \overline{\left\{ \bigcup_{k=0}^{\infty} S_{d_N}^k(w^\star, \cdots, w^\star) \right\}} = \overline{\left\{ \bigcup_{k=N}^{\infty} S_x^k(w^\star, \cdots, w^\star) \right\}}.$$

It immediately follows that

$$d^\star \in \bigcap_{N=0}^{\infty} \overline{\left\{ \bigcup_{k=N}^{\infty} S_x^k(w^\star, \cdots, w^\star) \right\}},$$

and hence condition A5 is equivalent to the condition that $d^\star$ is an ω-limit point of the system (10) for every initial condition (see [Saperstone, 1981]).

By replacing the function $F(\cdot,\cdot)$ used in (4) with $F(\cdot + d^\star, \cdot + w^\star) - d^\star$ and translating the state space $\mathbf{X}$ we may replace the constants $w^\star$ and $d^\star$ with zero. So, we henceforth assume that

$$0 \in \mathbf{X}, \quad w^\star = 0, \quad \text{and} \quad d^\star = 0.$$

We say the Markovian system (4) *satisfies condition* GA if 0 is globally attracting for the freely evolving system. That is:

GA. For each initial condition $x \in \mathbf{X}$,

$$\lim_{k\to\infty} d_k = \lim_{k\to\infty} S_x^k(0, \cdots, 0) = 0.$$

Hence if the disturbance sequence $\mathbf{w}$ is replaced by $(0, \cdots, 0, \cdots)$ in (4) then $\Phi_k \to 0$ as $k \to \infty$ for all initial conditions. It is easily verified that condition GA implies condition A5.

For example, the controlled random parameter $AR(p)$ system examined in [Meyn and Caines, 1987] satisfies condition GA when $\sigma_e^2 < 1$, and it is shown in [Meyn, 1989] that a linear system under nonlinear control satisfies condition GA under extremely general conditions.

The following result greatly simplifies the task of verifying weak stochastic controllability.

PROPOSITION 2.1. *Suppose conditions* A1–A5 *hold, and that* $0 \in \operatorname{supp} \mu_w$. *Then the following statements are equivalent*:

(i) $\mathbf{\Phi}$ *is w.s.c.*;

(ii) *the controllability matrix* $C_0^T(\lambda)$ *is full rank for some* $T \in \mathbb{Z}_+$ *and* $\lambda \in \mathcal{O}_w^T$;

(iii) *for open sets* $0 \in \mathcal{U}$ *and* $\mathcal{V}$, $T \in \mathbb{Z}_+$, *and a constant* $c > 0$,

$$P^T(x, A) \geqq c\mathbf{1}_{\mathcal{U}}(x)\mu^{\text{Leb}}\{A \cap \mathcal{V}\} \quad \textit{for all } x \in \mathbf{X} \textit{ and } A \in \mathcal{B}(\mathbf{X}).$$

The following corollary follows immediately.

COROLLARY 2.2. *If conditions* A1–A4 *and* GA *hold, and* $0 \in \operatorname{supp} \mu_w$ *then* $\mathbf{\Phi}$ *is w.s.c. if the pair* $(A_0(0), B_0(0))$ *is controllable in the usual sense.*

Proof of Corollary 2.2. Under the given assumptions 0 is a fixed point of the dynamical system (10). Hence if $(A, B) \triangleq (A_0(0), B_0(0))$ is controllable, it follows that

$$C_0^n(0) = [A^{n-1}B| \cdots |AB|B] \quad \text{is full rank.}$$

Since $0 \in \operatorname{supp} \mu_w = \bar{\mathcal{O}}_w$ and the matricial function $C_0^n(\cdot)$ is continuous, it follows that $C_0^n(\lambda)$ is full rank for some $\lambda \in \mathcal{O}_w$ sufficiently close to 0. Hence condition (ii) of Proposition 2.1 holds. □

Before proving Proposition 2.1 we must establish the following lemma.

LEMMA 2.1. *Suppose that the Markov chain* $\mathbf{\Phi}$ *is generated by the Markovian system* (4) *satisfying conditions* A1–A4, *and that* $0 \in \operatorname{supp} \mu_w$. *Then for each* $x \in \mathbf{X}$, *and every open set* $\mathcal{U} \subset \mathbf{X}$ *containing the origin,*

(i) *if condition* A5 *is satisfied then*

$$\sup_{k \geqq 0} P^k(x, \mathcal{U}) > 0;$$

(ii) *if condition* GA *is satisfied then*

$$P^k(x, \mathcal{U}) > 0 \quad \textit{for all sufficiently large } k \in \mathbb{Z}_+.$$

Proof. Fix $x \in \mathbf{X}$, and let $\mathcal{U}$ satisfy the hypotheses of the lemma. If condition A5 holds we may choose $k \in \mathbb{Z}_+$ such that

$$S_x^k(0, \cdots, 0) \in \mathcal{U}, \tag{12}$$

and by continuity there exists a $\delta > 0$ such that

$$S_x^k(z_1, \cdots, z_k) \in \mathcal{U},$$

for all $(z_1, \cdots, z_k) \in \{B_\delta(0)\}^k$, where $B_\delta(0)$ is the open rectangle of width δ centered at the origin. It follows that

$$\begin{aligned} P^k(x, \mathcal{U}) &> E[\mathbf{1}_{\|w_1\|<\delta} \cdots \mathbf{1}_{\|w_k\|<\delta}] \\ &= (\mu_w\{B_\delta(0)\})^k > 0. \end{aligned}$$

If condition GA holds then (12) is satisfied for all k sufficiently large, and so by the same argument as before, $P^k(x, \mathcal{U}) > 0$ for such k. □

Proof of Proposition 2.1. By Theorem 2.1, condition (i) of Proposition 2.1 implies condition (ii), which implies condition (iii). To complete the proof we will show that condition (iii) implies condition (i).

Suppose condition (iii) holds. Fix $x \in \mathbf{X}$ and choose an integer k for which $P^k(x, \mathscr{U}) > 0$. This is possible by condition (iii) and Lemma 2.1.

Then for any $A \in \mathscr{B}(\mathbf{X})$,

$$P^{k+T}(x, A) = \int P^k(x, dy) P^T(y, A) \geqq c P^k(x, \mathscr{U}) \mu^{\text{Leb}}\{A \cap \mathscr{V}\}.$$

Since $cP^k(x, \mathscr{U}) > 0$, we conclude that $\mathbf{1}\mathscr{V}\mu^{\text{Leb}} < P^{T+k}(x, \cdot)$, and hence $\mathbf{\Phi}$ is w.s.c. □

Remark. It is easy to see that condition A5 may be considerably relaxed. The following weaker condition (implied by asymptotic controllability—see [Sontag, 1983]) may be used to replace condition A5 in all of the results of this paper:

For each $x \in \mathbf{X}$ and $\varepsilon > 0$, there exists $N \in \mathbb{Z}_+$ and a deterministic sequence $\{w_k^\star : 1 \leqq k \leqq N\}$ lying in $\mathscr{O}_w^N$ such that

$$|S_x^N(w_1^\star \cdots w_N^\star)| < \varepsilon. \tag{13}$$

Hence conditions GA and A5 are useful because they imply that the *deterministic* system (4) (with the disturbance $\mathbf{w}$ considered as a deterministic input) is controllable to the origin.

2.2. Stability. Here we introduce some useful stability conditions, and then in Proposition 2.2 we give necessary and sufficient conditions for (4) to be bounded in probability on average.

DEFINITION. (i) The Markovian system (4) is called *bounded in probability* if for each deterministic initial condition $x \in \mathbf{X}$ and each $\varepsilon > 0$ there exists a compact subset $C \subset \mathbf{X}$ such that

$$\liminf_{k \to \infty} P^k(x, C) \geqq 1 - \varepsilon.$$

(ii) The Markovian system (4) is called *bounded in probability on average* if for each deterministic initial condition $x \in \mathbf{X}$ and each $\varepsilon > 0$ there exists a compact subset $C \subset \mathbf{X}$ such that

$$\liminf_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} P^k(x, C) \geqq 1 - \varepsilon.$$

(iii) The Markovian system (4) is called *uniformly bounded in probability* if for each $\varepsilon > 0$ there exists a compact subset $C \subset \mathbf{X}$ such that

$$\liminf_{k \to \infty} P^k(x, C) \geqq 1 - \varepsilon$$

for all $x \in \mathbf{X}$. □

For example, if $\mathbf{X} = \mathbb{R}^n$ and

$$\limsup_{k \to \infty} E_x[|\Phi_k|^2] < \infty \quad \text{for every } x \in \mathbf{X},$$

then $\mathbf{\Phi}$ is bounded in probability, and if for some fixed constant $J < \infty$,

$$\limsup_{k \to \infty} E_x[|\Phi_k|^2] \leqq J \quad \text{for every } x \in \mathbf{X},$$

then $\mathbf{\Phi}$ is uniformly bounded in probability.

In stochastic control theory we are often concerned with showing that the performance criteria

$$L_\infty \triangleq \limsup_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} y_k^2 + \rho u_k^2 \quad \text{and} \quad J_\infty \triangleq \limsup_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} E[y_k^2 + \rho u_k^2]$$

are uniformly bounded for all initial conditions. If this is the case, then it often follows that $\mathbf{\Phi}$ is bounded in probability on average but it does not necessarily follow that it is bounded in probability. However if (4) is w.s.c. we will find that the two notions of stability are equivalent. This is one of the main results to be presented below.

Here we state a necessary and sufficient condition for the second form of stability.

PROPOSITION 2.2. *The Markovian system* (4) *is bounded in probability on average if and only if for each deterministic* $x \in \mathbf{X}$ *there exists a nonempty, tight collection of invariant probabilities* Π_x *such that*

$$\frac{1}{N}\sum_{k=1}^{N} P^k(x, \cdot) \xrightarrow{\text{weakly}} \Pi_x. \tag{14}$$

Proof. Fix $x \in \mathbf{X}$, let $\mu_k = P^k(x, \cdot)$, and suppose (4) is bounded in probability on average. In this case the collection of probabilities $\{(1/N)\sum_{k=1}^{N} \mu_k,\ k \geqq 1\}$ is a precompact subset of the space of all probabilities on $\mathscr{B}(\mathbf{X})$, and it is a routine exercise to show that any weak limit point is invariant. For each initial condition $x \in \mathbf{X}$, the set of invariant probabilities obtained in this way is tight under the conditions of Proposition 2.2 because the set of limit points of a precompact set is compact. The proof of the converse is straightforward, and so we omit it. □

2.3. Main results. Here we state the main results of the paper. We begin by presenting some definitions from the theory of Markov chains on general state spaces.

Let $\Lambda : \mathbf{X} \times \mathscr{B}(\mathbf{X}) \to [0, 1]$ denote the function

$$\Lambda(x, A) \triangleq P_x\{\mathbf{\Phi} \text{ enters } A\} = P_x\left\{\bigcup_{k=0}^{\infty} \{\Phi_k \in A\}\right\}. \tag{15}$$

A set $A \in \mathscr{B}(\mathbf{X})$ is called *absorbing* if it is nonempty and $P(x, A) = 1$ for each $x \in A$. An absorbing set H is called a *Harris set* if there exists a probability μ such that $\Lambda(x, A) = 1$ for all $x \in H$ whenever $\mu\{A\} > 0$. If the state space $\mathbf{X}$ is a Harris set, then $\mathbf{\Phi}$ is called *Harris recurrent.*

If a Harris set H exists, then there exists a unique, up to constant multiples, invariant measure π that is supported on H. When π is finite, then H is called a positive Harris set, and in the case where $H = \mathbf{X}$, we call $\mathbf{\Phi}$ positive Harris recurrent.

We will only consider positive Harris sets in this paper, and to simplify the terminology we will henceforth drop the adjective "positive."

Harris sets are periodic: There exists a maximal integer λ called the period of H, and disjoint measurable sets $\{E_i : 1 \leqq i \leqq \lambda\}$ such that $H = \bigcup E_i$ and

$$P(x, E_{i+1}) = 1, \qquad x \in E_i \pmod{\lambda};$$

when $\lambda = 1$ we call H (or $\mathbf{\Phi}$ in the case where $\mathbf{X} = H$) *aperiodic.*

Our interest in Harris sets comes from the following result. Let $\|\cdot\|_{tv}$ denote the total variation norm, defined on the space of finite signed measures on $\mathscr{B}(\mathbf{X})$.

PROPOSITION 2.3. $\mathbf{\Phi}$ *is positive Harris recurrent if and only if there exists a unique invariant probability* π *and an integer* λ *such that the following limits hold*:

$$\lim_{N\to\infty} \frac{1}{N}\sum_{k=1}^{N} f(\Phi_k) = \int f\,d\pi \quad a.s.\ [P_x], \tag{16}$$

$$\lim_{k\to\infty} \left\| \frac{1}{\lambda}\sum_{i=1}^{\lambda} P^{k+i}(x, \cdot) - \pi \right\|_{tv} = 0 \tag{17}$$

for every initial condition $x \in \mathbf{X}$, *and every function* $f \in L^1(\mathbf{X}, \mathscr{B}(\mathbf{X}), \pi)$.

Proof. It is shown in [Athreya and Ney, 1980] that if $\mathbf{\Phi}$ is positive Harris recurrent with invariant probability π, then (16) holds for every initial distribution, and in fact a generalization of (16) holds even in the null recurrent case (when the invariant measure π is not finite).

Conversely, if (16) holds for every initial condition, then it is obvious that $\Lambda(x, A)=1$ for every set A of positive π-measure, and hence $\mathbf{\Phi}$ is positive Harris recurrent.

A proof of the equivalence between Harris recurrence and the limit (17) may be found in [Nummelin, 1984]. □

Stability, controllability, and ergodicity. The following result is a direct consequence of Theorem 8.2 and the corollary to Theorem 9.1 of [Orey, 1971]. The hypotheses of these theorems are satisfied since if $\mathbf{\Phi}$ is w.s.c. then for each $x \in \mathbf{X}$ the probability

$$\sum_{k=0}^{\infty} 2^{-k-1} P^k(x, \cdot)$$

is nonsingular with respect to Lebesgue measure.

We remark that Proposition 2.4 also follows from (9) and the Decomposition Theorem of [Tuominen and Tweedie, 1979] by constructing a suitable continuous component.

PROPOSITION 2.4. *Suppose conditions* A1–A4 *hold and that* $\mathbf{\Phi}$ *is w.s.c. Then there exists a countable (possibly empty) index set I, a collection of disjoint Harris sets* $\{H_i \colon i \in I\}$, *and corresponding invariant probabilities* $\{\pi^i \colon i \in I\}$. *Furthermore, there exists at least one finite invariant measure* μ *if and only if I is nonempty, and in this case it has the form* $\mu = \sum q_i \pi^i$ *for a summable sequence* $\{q_i \colon i \in I\} \subset \mathbb{R}_+$.

The following functions will be of great use in studying the asymptotic behavior of $\mathbf{\Phi}$ with arbitrary initial conditions. Recall from (15) that $\Lambda(x, A) = P_x\{\mathbf{\Phi} \text{ enters } A\}$. For $x \in \mathbf{X}$ and $i \in I$ let

$$\alpha_i(x) = \Lambda(x, H_i), \qquad \alpha(x) = \sum_{i \in I} \alpha_i(x). \tag{18}$$

Of course, if the index set I is empty, then we have $\alpha \equiv 0$.

Although there may be more than one Harris set associated with one of the invariant probabilities π^i, the definition above does not depend on the Harris set chosen. Since H_i and $\cup H_i$ are absorbing

$$\alpha_i(x) = \lim_{k \to \infty} P^k(x, H_i), \qquad \alpha(x) = \lim_{k \to \infty} P^k\left(x, \bigcup_{i \in I} H_i\right).$$

In the following result weak stochastic controllability is used to establish the equivalence of boundedness in probability and its averaged formulation, and the Strong Law of Large Numbers for Markov chains [Doob, 1953] is generalized to arbitrary initial conditions.

If $\alpha(x) > 0$ we define the invariant probability π_x by

$$\pi_x = \frac{1}{\alpha(x)} \sum_{i \in I} \alpha_i(x) \pi^i. \tag{19}$$

THEOREM 2.2. *Suppose that conditions* A1–A4 *hold. If* (4) *is w.s.c. then the following are equivalent*:

(i) (4) *is bounded in probability*;

(ii) (4) *is bounded in probability on average*;

(iii) $\alpha(x)=1$ *for all* $x\in\mathbf{X}$;

(iv) *for each* $x\in\mathbf{X}$ *and* $f\in L^1(\mathbf{X}^{\mathbb{Z}_+},\mathscr{B}(\mathbf{X}^{\mathbb{Z}_+}),P_{\pi_x})$ *there exists* $f_\infty\in L^1(\mathbf{X}^{\mathbb{Z}_+},\mathscr{B}(\mathbf{X}^{\mathbb{Z}_+}),P_x)$ such that

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} f(\Phi_k,\Phi_{k+1},\cdots)=f_\infty(\Phi_0,\Phi_1,\cdots)\quad a.s.\ [P_x];$$

(v) *for each* $x\in\mathbf{X}$,

$$\left\|\frac{1}{N}\sum_{k=1}^{N} P^k(x,\cdot)-\pi_x\right\|_{tv}\to 0$$

as $N\to\infty$, *uniformly for x in compact subsets of* $\mathbf{X}$.

The precise form of the random variable f_∞ used in (iv) is easy to guess. On the event $\{\Phi$ enters $H_i\}$ the limiting behavior of $(1/N)\sum_{k=1}^{N}f(\Phi_k,\Phi_{k+1},\cdots)$ is determined by π^i with probability one. Hence

$$f_\infty=\mathbf{1}_{\{\tau<\infty\}}\sum_{k\in I}\mathbf{1}\{\Phi_\tau\in H_k\}E_{\pi^k}[f(\Phi_0,\Phi_1,\cdots)],\tag{20}$$

where τ is the first entrance time into the set $H=\bigcup_{i\in I}H_i$:

$$\tau=\min\{k\in\mathbb{Z}_+:\Phi_k\in H\}.$$

The relevance of (iii) is greater than might first appear. To see how Theorem 2.2 may be applied to a specific control problem, consider the controlled system of [Meyn and Caines, 1987]. A direct stability proof for the Markovian state process of that system is difficult, but the existence of an invariant probability π is relatively easy to establish and it is then obvious that the state process enters the open set that supports π almost surely. Hence by Theorem 2.2 the system is bounded in probability and in particular (iv) and (v) hold for that example.

The following corollary immediately follows.

COROLLARY 2.3. *Suppose that* (4) *is w.s.c. Then* $\mathbf{\Phi}$ *is positive Harris recurrent if and only if it is bounded in probability and possesses exactly one invariant probability.*

We now strengthen the stability hypothesis on $\mathbf{\Phi}$ to obtain correspondingly stronger limit theorems.

THEOREM 2.3. *Suppose that conditions* A1–A4 *hold, and that* $\mathbf{\Phi}$ *is w.s.c. Then* $\mathbf{\Phi}$ *is uniformly bounded in probability if and only if it is bounded in probability and the index set I is finite. If this is the case, then there exists* $\lambda\in\mathbb{Z}_+$ *such that*

$$\left\|\frac{1}{\lambda}\sum_{k=1}^{\lambda} P^{k+i}(x,\cdot)-\pi_x\{\cdot\}\right\|_{tv}\to 0\quad as\ i\to\infty$$

uniformly for x in compact subsets of $\mathbf{X}$.

THEOREM 2.4. *Suppose that conditions* A1–A4 *hold,* $0\in\operatorname{supp}\mu_w$, *and* $\mathbf{\Phi}$ *is w.s.c. and bounded in probability. Then*

(i) *if* A5 *is satisfied then a unique invariant probability π exists, and hence* $\mathbf{\Phi}$ *is positive Harris recurrent*;

(ii) *if* GA *is satisfied then* $\mathbf{\Phi}$ *is aperiodic and positive Harris recurrent, and in this case* $P^k(x,A)\to\pi$ *as* $k\to\infty$ *uniformly for* $A\in\mathscr{B}(\mathbf{X})$ *and x in compact subsets of* $\mathbf{X}$.

The proofs of these results will be given in § 3. We now show how these results may be related to Lyapunov function theory.

On the properties of Lyapunov functions. One of the difficulties in deterministic as well as stochastic stability studies is that stability is a property of the entire process, viewed from time 0 to ∞, while any useful test should make use of only a finite amount of data, e.g., values of the process at specific (finite) times, or the relationship between values of the process at different finite times.

Lyapunov's second method or its stochastic generalizations [Kushner, 1967], [Has'minskiĭ, 1980] are among the most successful approaches to this problem that meet these requirements.

One specific example of a stochastic generalization of Lyapunov's second method may be found in the stability proof of [Goodwin, Ramadge, and Caines, 1981]. The methodology of that paper may be very roughly described as follows. A positive adapted stochastic process $\{V_k\}$ is constructed satisfying the super martingale property

$$E[V_{k+1} \mid \mathscr{F}_k] \leqq V_k, \tag{21}$$

where $\mathscr{F}_k$ is the σ-algebra generated by past and present values of $\mathbf{\Phi}$,

$$\mathscr{F}_k \triangleq \sigma\{\Phi_0, \cdots, \Phi_k\}, \tag{22}$$

and $\mathbf{\Phi}$ denotes a state process for the closed-loop system under consideration.

From the fact that $\{V_k\}$ is a convergent super martingale and other specific properties of this process we may deduce mean-square stability and, in some sense, optimality of the performance of the closed-loop system.

This approach will fail in the case of a w.s.c. Markovian system in many cases of interest. For example, if $V_k = V(\Phi_k)$ for a continuous function $V: \mathbf{X} \to \mathbb{R}_+$, then $\mathbf{\Phi}$ converges to a level set of the function V. If $\mathbf{\Phi}$ is bounded in probability and w.s.c., then by Theorem 2.2 the set of limit points of the sequence $\{\Phi_k : k \in \mathbb{Z}_+\}$ is equal to $\sum_{i \in I} \mathbf{1}_{\{\mathbf{\Phi} \text{ enters } S_i\}} \operatorname{supp} \pi^i$ almost surely. By weak stochastic controllability, the support of any invariant probability has nonempty interior, which implies that the function V must be flat over suitably large regions of the state space. For example, this rules out real-analytic functions. Since in many examples the function V is in fact a rational function of its arguments, it may be seen that weak stochastic controllability rules out a large class of test functions of the form (21).

However the following alternative stability test has already been of great use in a number of examples (see [Guo and Meyn, 1989], [Meyn and Guo, 1990] and [Meyn, 1989]) and has great theoretical potential. As before, let $\{V_k\}$ be a positive adapted stochastic process, and suppose that for some $0 < \lambda < 1$, $L > 0$,

$$E[V_{k+1} \mid \mathscr{F}_k] \leqq \lambda V_k + L, \tag{23}$$

which in the degenerate case $\lambda = 1$, $L = 0$ becomes (21). Under this condition and certain technical assumptions that include weak stochastic controllability, it may be shown that the underlying distributions of $\mathbf{\Phi}$ converge to an invariant probability at a geometric rate, and that the central limit theorem holds for functions whose square is dominated by $\{V_k\}$. Some of these results may be found in [Meyn and Tweedie, 1990a,b], and in [Meyn, 1989] where processes that are not w.s.c. are also considered.

3. Proofs of the main results.

3.1. Proof of Theorem 2.1. To prove Theorem 2.1 we will need the following lemma. For $x \in \mathbb{R}^m$ and $\varepsilon > 0$, let $B_\varepsilon(x)$ denote the open rectangle

$$B_\varepsilon(x) \triangleq \{y \in \mathbb{R}^m : |x_i - y_i| < \varepsilon \text{ for } 1 \leqq i \leqq m\}.$$

LEMMA 3.0. *Let $\mathscr{W}_1 \subset \mathbb{R}^n$, $\mathscr{X}_1 \subset \mathbb{R}^m$, and $\mathscr{Y}_1 \subset \mathbb{R}^n$ be open and suppose $G: \mathscr{W}_1 \times \mathscr{X}_1 \times \mathscr{Y}_1 \to \mathbb{R}^n$, $(w, x, y) \to z$, is smooth, and that the matrix $\partial G / \partial y$ is full rank at some $(w_0, x_0, y_0) \in \mathscr{W}_1 \times \mathscr{X}_1 \times \mathscr{Y}_1$. Then*

(i) *there exists an open set*

$$\mathscr{W} \times \mathscr{X} \times \mathscr{Y} \subset \mathscr{W}_1 \times \mathscr{X}_1 \times \mathscr{Y}_1 \quad \textit{containing } (w_0, x_0, y_0)$$

such that the measure $\nu(w, \cdot)$ defined for $A \in \mathcal{B}(\mathbb{R}^n)$ and $w \in \mathcal{W}$ by

$$\nu(w, A) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbf{1}_{G(w,x,y) \in A} \, dx \, dy$$

is equivalent to Lebesgue measure on an open set $\mathcal{R}_w$;

(ii) *there exists $c > 0$ and open sets $\mathcal{U}$ and $\mathcal{V}$ containing w_0 and $G(w_0, x_0, y_0)$, respectively, such that*

$$\nu(w, \cdot) \geqq c\mathbf{1}_{\mathcal{U}}(w)(\mathbf{1}_{\mathcal{V}} \mu^{\mathrm{Leb}}) \quad \textit{for all } w.$$

Proof of Lemma 3.0. Consider the function $G^\star: \mathcal{W}_1 \times \mathcal{X}_1 \times \mathcal{Y}_1 \to \mathbb{R}^{n+m+n}$ defined for $(w, x, y) \in \mathcal{W}_1 \times \mathcal{X}_1 \times \mathcal{Y}_1$ by

$$G^\star(w, x, y) \triangleq \begin{pmatrix} w \\ x \\ G(w, x, y) \end{pmatrix}.$$

Under the conditions of Lemma 3.0 the function $G^\star$ is smooth, and its derivative is full rank at (w_0, x_0, y_0). By the implicit function theorem there exist open sets

$$\mathcal{R} \subset \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \quad \text{and} \quad \mathcal{W} \times \mathcal{X} \times \mathcal{Y} \subset \mathcal{W}_1 \times \mathcal{X}_1 \times \mathcal{Y}_1$$

with $(w_0, x_0, y_0) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Y}$, and a smooth function $H^\star: \mathcal{R} \to \mathcal{W} \times \mathcal{X} \times \mathcal{Y}$ such that

$$\mathcal{R} = \left\{ \begin{pmatrix} w \\ x \\ G(w, x, y) \end{pmatrix} : w \in \mathcal{W},\ x \in \mathcal{X},\ y \in \mathcal{Y} \right\}$$

and

$$H^\star(G^\star(w, x, y)) = (w, x, y)$$

for $(w, x, y) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Y}$. Applying a projection to the function $H^\star$ we may find a smooth function $H: \mathcal{R} \to \mathcal{Y}$ for which

$$H(w, x, G(w, x, y)) = H(G^\star(w, x, y)) = y$$

for $(w, x, y) \in \mathcal{W} \times \mathcal{X} \times \mathcal{Y}$.

Since H is smooth and

$$\left[\frac{\partial H}{\partial z}\right] = \left[\frac{\partial G}{\partial y}\right]^{-1}$$

is full rank, we may assume that

$$\left| \det \left[\frac{\partial H}{\partial z}\right] \right| \geqq h_0 \tag{24}$$

for all $(w, x, z) \in \mathcal{R}$.

We now construct a density for the kernel

$$\nu(w, A) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbf{1}_{G(w,x,y) \in A} \, dx \, dy, \qquad w \in \mathcal{W}, \quad A \in \mathcal{B}(\mathbb{R}^n)$$

by a change of variables and Fubini's theorem:

$$\begin{aligned} \nu(w, A) &= \int_{\mathcal{X}} \left\{ \int_{G(w,x,\mathcal{Y})} \mathbf{1}_A(z) \left| \det \frac{\partial H}{\partial z} \right| dz \right\} dx \\ &= \int_{\mathbb{R}^n} \mathbf{1}_A(z) \left\{ \int_{\mathbb{R}^m} \mathbf{1}_{\mathcal{R}}(w, x, z) \left| \det \frac{\partial H}{\partial z} \right| dx \right\} dz. \end{aligned}$$

Hence for fixed $w \in \mathcal{W}$, the function $p : \mathbb{R}^n \to \mathbb{R}_+$ defined for $z \in \mathbb{R}^n$ by

$$p(w, z) \triangleq \int_{\mathbb{R}^m} \mathbf{1}_{\mathcal{R}}(w, x, z) \left| \det \frac{\partial H}{\partial z} \right| dx \tag{25}$$

is a density for $\nu(w, \cdot)$.

Fix $w \in \mathcal{W}$, and let $\mathcal{R}_w \subset \mathbb{R}^n$ denote the open set

$$\mathcal{R}_w \triangleq \{z \colon (w, x, z) \in \mathcal{R} \text{ for some } x \in \mathcal{X}\}.$$

Then by (24) and (25) it is easy to show that $p(w, \cdot)$ is strictly positive on $\mathcal{R}_w$ and zero elsewhere. Hence $\nu(w, \cdot)$ is equivalent to Lebesgue measure on $\mathcal{R}_w$ and this proves (i).

Since $\mathcal{R}$ is an open neighborhood of (w_0, x_0, z_0), there exists a nonempty open rectangle

$$(w_0, x_0, z_0) \in \mathcal{W}_0 \times \mathcal{X}_0 \times \mathcal{Z}_0 \subset \mathcal{R}.$$

Hence by (24) and (25),

$$\begin{aligned} p(w, z) &\geqq h_0 \int_{\mathbb{R}^m} \mathbf{1}_{\mathcal{W}_0}(w) \mathbf{1}_{\mathcal{X}_0}(x) \mathbf{1}_{\mathcal{Z}_0}(z) \, dx \\ &= h_0 \mu^{\text{Leb}}\{\mathcal{X}_0\} \mathbf{1}_{\mathcal{W}_0}(w) \mathbf{1}_{\mathcal{Z}_0}(z). \end{aligned}$$

It follows that

$$\nu(w, A) \geqq (h_0 \mu^{\text{Leb}}\{\mathcal{X}_0\}) \mathbf{1}_{\mathcal{W}_0}(w) (\mathbf{1}_{\mathcal{Z}_0} \mu^{\text{Leb}})\{A\} \qquad w \in \mathbb{R}^n, \quad A \in \mathcal{B}(\mathbb{R}^n)$$

and this establishes (ii). □

Proof of Theorem 2.1. Fix a point $x \in \mathbf{X}$, $T \in \mathbb{Z}_+$, and let $Z_x = Z_x(T)$ denote the closed subset of $\mathbb{R}^{pT}$,

$$Z_x \triangleq \{\lambda \in \mathbb{R}^{pT} \colon \operatorname{rank} C_x^T(\lambda) \leqq n - 1\}.$$

Proof of (ii). Suppose that rank $C_x^T(\lambda) < n$ for all $\lambda \in \mathcal{O}_w^T$. By Sard's theorem, the set $S = \{S_x^T(\lambda) \colon \lambda \in \mathcal{O}_w^T\}$ has zero Lebesgue measure. Hence in this case $P^T(x, \cdot)$ is supported on S, and it is therefore singular with respect to Lebesgue measure. This shows that the rank condition (7) is necessary.

To prove sufficiency suppose (7) is satisfied for some $\lambda_0 \in \mathcal{O}_w^T \backslash Z_x$. Let p_w denote the density for μ_w, and p the density defined for $(z_1, \cdots, z_T) \in \mathbb{R}^{Tp}$ by

$$p(z_1, \cdots, z_T) \triangleq p_w(z_1) \cdots p_w(z_T).$$

Since $\lambda_0 \in \mathcal{O}_w^T$ and p_w is lower semicontinuous we may find $p_0 > 0$ and an open rectangle $B_\delta(\lambda_0)$ such that

$$p(\lambda) \geqq p_0 \quad \text{for all } \lambda \in B_\delta(\lambda_0).$$

Hence

$$\begin{aligned} P^T(y, \cdot) &= \int \mathbf{1}_{S_y^T \in} \cdot p(\lambda) \, d\lambda \\ &\geqq p_0 \int_{B_\delta(\lambda_0)} \mathbf{1}_{S_y^T \in \cdot} \, d\lambda \quad \text{for all } y \in \mathbf{X}. \end{aligned} \tag{26}$$

Using the rank condition on C_x^T we may find integers $\{i_1, \cdots, i_n\}$ for which

$$\det \left[\frac{\partial S_x^T}{\partial \lambda_{i_1}} \Big| \cdots \Big| \frac{\partial S_x^T}{\partial \lambda_{i_n}} \right]_{\lambda_0} \neq 0.$$

This allows us to apply Lemma 3.0. By defining the function G in terms of S_x^T appropriately, all the conditions of that lemma are satisfied, and by reducing the size of δ if necessary, open sets $\mathcal{U}_x^{\lambda_0}$ and $\mathcal{V}_x^{\lambda_0}$ containing x and $S_x^{\lambda_0}$, respectively, an open set $\mathcal{R}_{\lambda_0}$, and a constant $c>0$ exist such that

$$(27)\qquad \begin{aligned} &\text{(i)} & &\int_{B_\delta(\lambda_0)} \mathbf{1}\{S_x^T\in\cdot\}\,d\lambda \approx \mathbf{1}_{\mathcal{R}_{\lambda_0}}\mu^{\mathrm{Leb}} \\ &\text{(ii)} & &\int_{B_\delta(\lambda_0)} \mathbf{1}\{S_x^T\in\cdot\}\,d\lambda \geqq c\mathbf{1}_{\mathcal{U}_x^{\lambda_0}}(y)(\mathbf{1}_{\mathcal{V}_x^{\lambda_0}}\mu^{\mathrm{Leb}}) \quad \text{for all } y\in\mathbf{X}, \end{aligned}$$

and $B_\delta(\lambda_0)\subset\mathcal{O}_w^T\backslash Z_x$. Combining (26) and (27)(i) shows that $P^T(x,\cdot)\succ\mathbf{1}_{\mathcal{R}_{\lambda_0}}\mu^{\mathrm{Leb}}$ and this proves (ii) of the theorem.

To prove (iii) observe that under the given hypotheses (26) and (27) hold and hence

$$P^T(y,\cdot)\geqq cp_0\mathbf{1}_{\mathcal{U}_x^{\lambda_0}}(y)(\mathbf{1}_{\mathcal{V}_x^{\lambda_0}}\mu^{\mathrm{Leb}})$$

for all $y\in\mathbf{X}$.

Proof of (i). Suppose that $\mu^{\mathrm{Leb}}\{Z_x\cap\mathcal{O}_w^T\}>0$, where Z_x is the set on which the matrix C_x^T is not full rank. By Sard's theorem the set $S=\{S_x^T(\lambda)\colon \lambda\in\mathcal{O}_w^T\cap Z_x\}$ has zero Lebesgue measure. Furthermore,

$$P^T(x,S)\geqq\int_{Z_x}\mathbf{1}\{S_s^T\in S\}\,d\mu_w^T=\mu_w^T\{Z_x\}>0,$$

and it follows that $P^T(x,\cdot)$ is not absolutely continuous with respect to Lebesgue measure. This shows that the rank condition (6) is necessary.

Under (6) and proceeding as above we may find for each $\lambda\in\mathcal{O}_w^T\backslash Z_x$ an open set $\mathcal{R}_\lambda$, and $\delta>0$ such that (27)(i) holds. We may assume that δ is so small that for some $p_0>0$,

$$p(\eta)>p_0 \quad \text{for all } \eta\in B_\delta(\lambda) \quad \text{and} \quad B_\delta(\lambda)\subset\mathcal{O}_w^T\backslash Z_x,$$

where the constant p_0 may depend on λ.

Define the set $\mathcal{O}_x$ by

$$\mathcal{O}_x\triangleq\{S_x^T(\lambda)\colon \lambda\in\mathcal{O}_w^T\backslash Z_x\}.$$

This set is open since it may also be written

$$\mathcal{O}_x=\bigcup_{\lambda\in\mathcal{O}_w^T\backslash Z_x}\mathcal{R}_\lambda.$$

The union of a family of open sets in $\mathbb{R}^N$ is equal to the union of some countable subfamily, and hence for constants δ_i, and open sets $\mathcal{R}_i$ we have

$$\mathcal{O}_x=\bigcup_{i=0}^\infty\mathcal{R}_i,\qquad \mathcal{O}_w^T\backslash Z_x=\bigcup_{i=0}^\infty B_{\delta_i},$$

with (26) and (27)(i) holding for each $i\in\mathbb{Z}_+$.

Hence $\mathbf{1}_{\mathcal{R}_i}\mu^{\mathrm{Leb}}\prec P^T(x,\cdot)$ for all $i\in\mathbb{Z}_+$, and it follows that $\mathbf{1}_{\mathcal{O}_x}\mu^{\mathrm{Leb}}\prec P^T(x,\cdot)$.

On the other hand,

$$\begin{aligned} P^T(x,\cdot)&\leqq\sum_{i=0}^\infty\int_{B_{\delta_i}}\mathbf{1}\{S_x^T\in\cdot\}p(\lambda)\,d\lambda \\ &\approx\sum_{i=0}^\infty\int_{B_{\delta_i}}\mathbf{1}\{S_x^T\in\cdot\}\,d\lambda \end{aligned}$$

since p is strictly positive on B_{δ_i}

$$\approx \sum_{i=0}^{\infty} \mathbf{1}_{\mathcal{R}_i} \mu^{\text{Leb}}$$

$$\approx \mathbf{1}_{\mathcal{O}_x} \mu^{\text{Leb}}.$$

Thus $P^T(x, \cdot) \approx \mathbf{1}_{\mathcal{O}_x} \mu^{\text{Leb}}$. □

3.2. Some technical lemmas. In this section we describe basic topological properties of the state space of locally and weakly stochastically controllable systems. These results will be used to establish limit theorems for the state process and its underlying distributions, and then the proofs of the main results of the paper in § 3.3.

Throughout the remainder of this paper we assume that conditions A1–A4 *hold and that* $\mathbf{\Phi}$ *is w.s.c.*

Let S_i denote the support of π^i, $\mathcal{U}_x$ and $\mathcal{V}_x$ will denote open sets satisfying

$$(28) \qquad P^T(y, A) \geqq c_x \mu^{\text{Leb}}\{A \cap \mathcal{V}_x\} \qquad A \in \mathcal{B}(\mathbf{X}), \quad y \in \mathcal{U}_x,$$

with $x \in \mathcal{U}_x$, $c_x > 0$, and $\mathcal{U}_i$ and $\mathcal{V}_i$ will denote some particular choice of $\mathcal{U}_x$ and $\mathcal{V}_x$ with $x \in S_i$. When I is nonempty, the existence of these sets is guaranteed by Theorem 2.1.

LEMMA 3.1. *If A is closed and absorbing then $\mathcal{V}_x \subset A$ for all $x \in A$.*

Proof. Let $x \in A$. Since A is absorbing, $P(x, A) = 1$, and since A is closed it follows that

$$\operatorname{supp} P^T(x, \cdot) \subset A.$$

By weak stochastic controllability and (28) we have $P^T(x, \cdot) \geqq c_x(\mathbf{1}_{\mathcal{V}_x} \mu^{\text{Leb}})\{\cdot\}$, and it follows that $\mathcal{V}_x \subset \operatorname{supp} P^T(x, \cdot)$. This shows that $\mathcal{V}_x \subset A$. □

LEMMA 3.2. *For each $i \in I$, the set, $\Lambda_i^{-1}\{0\} \triangleq \{y \in \mathbf{X}: \Lambda(y, \mathcal{U}_i) = 0\}$ is closed and absorbing.*

Proof. Since $\mathcal{U}_i$ is open, the function $\Lambda(\cdot, \mathcal{U}_i)$ is lower semicontinuous [Cogburn, 1975] and it follows that $\Lambda_i^{-1}\{0\}$ is closed.

It is a standard fact and easy to prove that $\Lambda_i^{-1}\{0\}$ is absorbing [Nummelin, 1984]. □

LEMMA 3.3. *For each $i \in I$ and $x \in \mathbf{X}$,*

(i) $\mathbf{1}_{\mathcal{V}_i} \mu^{\text{Leb}} < \pi^i$;

(ii) *$\mathcal{U}_x \cap S_j = \emptyset$ for all but possibly one $j \in I$;*

(iii) *$S_i \cap S_j = \emptyset$ for all $j \neq i$.*

Proof. Result (i) follows from the fact that $\mathcal{U}_i$ is a small set of positive π^i-measure [Nummelin, 1984] (see the discussion below (9)).

Result (ii) also follows from (i) since if for some $x \in \mathbf{X}$,

$$\mathcal{U}_x \cap S_i \neq \emptyset \quad \text{and} \quad \mathcal{U}_x \cap S_j \neq \emptyset \quad \text{then } \mathbf{1}_{\mathcal{V}_x} \mu^{\text{Leb}} < \pi^i, \pi^j.$$

This is impossible since π^i and π^j are mutually singular.

Finally, by (ii) and since $x \in \mathcal{U}_x$, if $i \neq j$,

$$S_j \subset \left\{ \bigcup_{x \in S_i} \mathcal{U}_x \right\}^c \subset S_i^c,$$

and this proves (iii). □

LEMMA 3.4. *For each compact set $K \subset \mathbf{X}$, $S_i \cap K = \emptyset$ for all but a finite number of integers $i \in I$.*

Proof. The collection of open sets $\{\mathcal{U}_x : x \in K\}$ cover K. Let $\{\mathcal{U}^1, \cdots, \mathcal{U}^N\}$ be a finite subcover. By Lemma 3.3 each $\mathcal{U}^i$ intersects at most one of the sets $\{S_i : i \in I\}$ and so no more than N of these sets intersect K. □

LEMMA 3.5. *For each $i \in I$, S_i does not contain any proper closed absorbing subsets.*

Proof. First of all, if A is an absorbing subset of S_i, then the function $\mathbf{1}_{A^c}$ is super harmonic. That is, $\int P(x, dy)\mathbf{1}_{A^c}(y) \leqq \mathbf{1}_{A^c}(x)$ for all $x \in \mathbf{X}$. By Proposition 3.13 of [Nummelin, 1984] it follows that $\pi^i\{A\}$ is equal to zero or one.

If A is also closed then by Lemma 3.1, $\mathcal{V}_x \subset A$ for every $x \in A$ and hence

$$\pi^i\{A\} \geqq \pi^i\{\mathcal{V}_x\} > 0,$$

which shows that in fact $\pi^i\{A\} = 1$. Since S_i is the smallest closed set that has full π_i-measure, it follows that $S_i = A$. □

LEMMA 3.6. *For each $x \in \mathbf{X}$,*

$$\{\Phi \textit{ enters } H_i\} = \{\Phi \in \mathcal{U}_i \text{ i.o.}\} = \{\lim_{k\to\infty} \Lambda(\Phi_k, \mathcal{U}_i) = 1\}$$

modulo P_x-null sets.

Proof. By a result of [Orey, 1971],

$$\lim_{k\to\infty} \Lambda(\Phi_k, \mathcal{U}_i) = \mathbf{1}_{\{\Phi \in \mathcal{U}_i \text{ i.o.}\}} \quad \text{a.s. } [P_x] \tag{29}$$

and so only the first equality requires verification. However this follows from (9), which expresses the fact that $\mathcal{U}_i$ is a small set of positive π^i-measure [Nummelin, 1984]. □

3.3. Asymptotic behavior. We now investigate the sample path properties and the asymptotic behavior of the underlying distributions of (4) under the weak stochastic controllability condition.

The following two results concern the asymptotic behavior of the Markov chain Φ restricted to one of the absorbing sets S_i. They are interesting in themselves and will also be useful for establishing limit theorems for arbitrary initial conditions.

If a set $A \in \mathcal{B}(\mathbf{X})$ is absorbing then for every initial condition $x \in A$, almost every $[P_x]$ sample path lies in $A^{\mathbb{Z}_+}$. Hence we can replace the state space $\mathbf{X}$ with A if we are only interested in such initial conditions.

The Markov chain Φ restricted to an absorbing set A is called *positive recurrent* if a Harris set $H \subset A$ exists and for each $x \in A$,

$$P_x\{\Phi \text{ enters } H\} > 0.$$

PROPOSITION 3.1. *Suppose conditions* A1–A4 *hold and Φ is weakly stochastically controllable. Then for each $i \in I$, the restriction of Φ to S_i is positive recurrent. If in addition Φ is bounded in probability, then S_i is a Harris set.*

Proof. Let $\mathcal{U}_i$, $\mathcal{V}_i$, and $\Lambda_i^{-1}\{0\}$ be defined as in § 3.2, and observe that by Lemma 3.2, $\Lambda_i^{-1}\{0\} \cap S_i$ is closed and absorbing. Since $\pi^i\{\mathcal{U}_i\} > 0$ it follows that

$$P_x\{\Phi \text{ enters } \mathcal{U}_i\} = 1 \quad \text{for all } x \in H_i.$$

Hence $\Lambda_i^{-1}\{0\} \subset H_i^c$ and it follows that $\pi^i\{\Lambda_i^{-1}\{0\}\} = 0$. So by Lemma 3.5, $\Lambda_i^{-1}\{0\} \cap S_i = \phi$. This implies that for every $x \in S_i$,

$$P_x\{\Phi \text{ enters } \mathcal{U}_i\} > 0. \tag{30}$$

We have, since $\mathbf{1}_{\mathcal{V}_i} < \pi^i$,

$$P_x\{\Phi \text{ enters } H_i\} \geqq \sup_{k\in\mathbb{Z}_+} P^{k+T}(x, H_i) \geqq \sup_{k\in\mathbb{Z}_+} cP^k(x, \mathcal{U}_i)\mu^{\text{Leb}}\{\mathcal{V}_i\}, \tag{31}$$

and combining (30) and (31) gives

$$P_x\{\mathbf{\Phi} \text{ enters } H_i\} > 0 \quad \text{for all } x \in S_i,$$

which shows that $\mathbf{\Phi}$ restricted to S_i is positive recurrent.

Suppose now that $\mathbf{\Phi}$ is bounded in probability. By Lemma 3.6, $\mathbf{\Phi}$ restricted to S_i is positive Harris recurrent if and only if

$$\lim_{k\to\infty} \Lambda(\Phi_k, \mathcal{U}_i) = 1 \quad \text{a.s. } [P_x] \tag{32}$$

for all $x \in S_i$.

Let $C \subset S_i$ be compact. By (30) the open sets

$$\mathcal{W}_k = \{x \in \mathbf{X} \colon \Lambda(x, \mathcal{U}_i) > 1/k\} \qquad k \geqq 1$$

form an open cover of C, and by compactness it follows that there exists $k_0 \in \mathbb{Z}_+$ such that

$$\Lambda(x, \mathcal{U}_i) > 1/k_0 \quad \text{for all } x \in C. \tag{33}$$

If the limit in (29) is zero then $\mathbf{\Phi}$ enters C only a finite number of times by (33). Since C is arbitrary, $\mathbf{\Phi}$ eventually leaves every compact set in this case. By stability, this can only happen on a P_x-null set and hence (32) holds. □

LEMMA 3.7. *Suppose conditions* A1–A4 *hold and that* $\mathbf{\Phi}$ *is l.s.c. Then for each* $x \in S_i$,

$$P^T(x, \cdot) < \pi^i,$$

where $T = T(x)$ *is the integer used in the definition of local stochastic controllability.*

Proof. Let $T = T(x)$, $\mathcal{O}_x$, Z_x, $\mathcal{V}_x^\lambda$, and $\mathcal{U}_x^\lambda$ be as in the proof of Theorem 2.1, and recall from Theorem 2.1 that

$$\mathcal{O}_x \triangleq \bigcup_{\lambda \in \mathcal{O}_w^T \setminus Z_x} S_x^T(\lambda) = \bigcup_{\lambda \in \mathcal{O}_w^T \setminus Z_x} \mathcal{V}_x^\lambda. \tag{34}$$

By Lemma 3.3(i),

$$\mathbf{1}_{\mathcal{V}_x^\lambda} \mu^{\text{Leb}} < \pi^i \quad \text{for all } x \in S_i \quad \text{and} \quad \lambda \in \mathcal{O}_w^T \setminus Z_x.$$

Since the union of a family of open sets in $\mathbf{X}$ is equal to the union of a countable subfamily, it follows from this and (34) that

$$P^T(x, \cdot) \approx \mathbf{1}_{\mathcal{O}_x} \mu^{\text{Leb}} < \pi^i. \qquad □$$

The following result follows immediately.

PROPOSITION 3.2. *If conditions* A1–A4 *hold and* $\mathbf{\Phi}$ *is l.s.c. then* S_i *is a Harris set for each* $i \in I$.

Proof. By Lemma 3.7 and the fact that $\pi^i\{H_i\} = 1$, it follows that $P^T(x, H_i) = 1$ for all $x \in S_i$, which by conditioning at time T and using the Markov property implies the desired result. □

We now show how the underlying distributions of w.s.c. Markovian systems exhibit asymptotically periodic behavior.

PROPOSITION 3.3. *If* (4) *is w.s.c. then for each initial condition* $x \in \mathbf{X}$, *the resulting trajectory* $\{\mu_k = P^k(x, \cdot) \colon k \in \mathbb{Z}_+\}$ *may be written*

$$\mu_k = n_k + \sum_{i \in I} \alpha_i(x) \mu_k^i, \tag{35}$$

where $\{n_k \colon k \in \mathbb{Z}_+\}$ *is a sequence of subprobabilities for which*

$$\frac{1}{N} \sum_{k=1}^{N} n_k \xrightarrow{\textit{vaguely}} 0, \tag{36}$$

and for each $i \in I$, the sequence of subprobabilities $\{\mu_k^i\colon k \in \mathbb{Z}_+\}$ converges in total variation norm to a periodic orbit; that is, there exists a periodic orbit $\{\gamma_k^i\colon k \in \mathbb{Z}_+\}$ such that

$$\lim_{k\to\infty}\left(\sup_{A\in\mathscr{B}(\mathbf{X})}|\mu_k^i\{A\}-\gamma_k^i\{A\}|\right)=0.$$

Proof. The proof of Proposition 3.3 will be completed in two steps.

Step 1. We show that for each $i \in I$, the sequence of subprobabilities

$$\{\mu_k^i \triangleq (1/\alpha_i(x))\mathbf{1}_{H_i}\mu_k\colon k\in\mathbb{Z}_+\}$$

converges to a periodic orbit whenever $\alpha_i(x) \neq 0$.

Fix $i \in I$. The Harris set H_i may be written as a disjoint union

$$H_i = \bigcup_{j=1}^{\lambda} E_j \cup \Delta, \quad \text{where } \pi^i\{\Delta\}=0,$$

and the cycle $\{E_j\colon 1 \leqq j \leqq \lambda\}$ has the invariance properties

$$P(y, E_{j+1}) = 1 \text{ for } y \in E_j \text{ (mod } \lambda),$$

and

$$P^\lambda(y, E_j) = 1 \text{ for } y \in E_j.$$

So, defining $\beta_k \in \mathbb{R}_+$ by $\beta_k \triangleq \lim_{n\to\infty} P^{n\lambda}(x, E_k)$ we have $\alpha_i(x) = \sum_{k=1}^{\lambda} \beta_k$.

If we let $d_k = \lambda \mathbf{1}_{E_k}\pi^i$,

$$\gamma_0 = \frac{1}{\alpha_i(x)}\sum_{k=1}^{\lambda}\beta_k d_k, \quad \text{and} \quad \gamma_k = \gamma_0 P^k \tag{37}$$

then $\gamma = \{\gamma_k\colon k \in \mathbb{Z}_+\}$ is a periodic orbit.

Since E_j is an aperiodic Harris set for the λ-step Markov chain $\{\Phi_{k\lambda}\colon k \in \mathbb{Z}_+\}$ (Proposition 3.14 of [Nummelin, 1984]) we have by Proposition 2.3

$$\lim_{k\to\infty}\sup_{B\in\mathscr{B}(\mathbf{X})}|\mathbf{1}_{E_j}\mu_{k\lambda}\{B\}-\beta_j d_j\{B\}|=0.$$

It follows that

$$\lim_{k\to\infty}\sup_{B\in\mathscr{B}(\mathbf{X})}\left|\frac{1}{\alpha_i(x)}\mathbf{1}_{H_i}\mu_{k\lambda}\{B\}-\gamma_0\{B\}\right|=0,$$

and further that for each $j \in \mathbb{Z}_+$,

$$\lim_{k\to\infty}\sup_{B\in\mathscr{B}(\mathbf{X})}\left|\frac{1}{\alpha_i(x)}\mathbf{1}_{H_i}\mu_{k\lambda+j}\{B\}-\gamma_j\{B\}\right|=0.$$

Hence $\{\mu_k^i\colon k \in \mathbb{Z}_+\}$ converges in total variation norm to the periodic orbit γ.

Step 2. We are left to show that with $H = \bigcup_{i\in I} H_i$, and $n_k \triangleq \mathbf{1}_{H^c}\mu_k$,

$$\frac{1}{N}\sum_{k=1}^{N} n_k \xrightarrow{\text{vaguely}} 0. \tag{38}$$

The set of all subprobabilities on $\mathscr{B}(\mathbf{X})$ is sequentially compact with respect to vague convergence. Hence (38) will hold if every vague limit point of the set of subprobabilities

$$\left\{\frac{1}{N}\sum_{k=1}^{N} n_k\colon N \in \mathbb{Z}_+\right\} \tag{39}$$

is zero. Let n_∞ be a vague limit point of (39) so that for some subsequence $\{N_i\}$ of $\mathbb{Z}_+$,

$$\lim_{i\to\infty}\frac{1}{N_i}\sum_{k=1}^{N_i}\int f\,dn_k \triangleq \lim_{i\to\infty}\frac{1}{N_i}\sum_{k=1}^{N_i} E_x[f(\Phi_k)\mathbf{1}_{H^c}(\Phi_k)] \tag{40}$$
$$= \int f\,dn_\infty$$

whenever $f\in C_c$. By choosing a further subsequence if necessary, we may assume that a subprobability μ_∞ exists such that

$$\frac{1}{N_i}\sum_{k=1}^{N_i}\mu_k \xrightarrow{\text{vaguely}} \mu_\infty \quad \text{as } i\to\infty.$$

It is easy to show that $\mu_\infty \geqq n_\infty$ (it is obvious that $\int f\,dn_\infty \leqq \int f\,d\mu_\infty$ for all $f\in C_c$) and by a proof similar to the result in [Foguel, 1969] we may show that μ_∞ is invariant.

Since every finite invariant measure is supported on $\cup\, S_i$, it follows that

$$n_\infty\left\{\bigcap_{i\in I} S_i^c\right\} \leqq \mu_\infty\left\{\bigcap_{i\in I} S_i^c\right\} = 0. \tag{41}$$

We now show that $n_\infty\{\cup_{i\in I} S_i\}=0$. From (41) it will follow that $n_\infty = 0$, completing the proof.

Let $i\in I$, $x\in S_i$, and $\mathcal{U}_x$ as before. If $\mathbf{\Phi}$ does not enter H_i then from Lemma 3.6 it follows that $\mathbf{\Phi}$ enters $\mathcal{U}_x$ finitely often and so,

$$\lim_{k\to\infty} \mathbf{1}_{H^c}(\Phi_k) - \mathbf{1}_{H^c}(\Phi_k)\mathbf{1}_{\mathcal{U}_x^c}(\Phi_k) = 0 \quad \text{a.s. } [P_x]. \tag{42}$$

Let $f\in C_c$ be any function which vanishes on $\mathcal{U}_x^c$. By (40),

$$\int f\,dn_\infty = \lim_{i\to\infty}\frac{1}{N_i}\sum_{k=1}^{N_i} E_x[f(\Phi_k)\mathbf{1}_{H^c}(\Phi_k)]$$
$$= \lim_{i\to\infty}\frac{1}{N_i}\sum_{k=1}^{N_i} E_x[f(\Phi_k)\mathbf{1}_{\mathcal{U}_x^c}(\Phi_k)\mathbf{1}_{H^c}(\Phi_k)]$$

by (42) and the Dominated Convergence Theorem

$$=0.$$

Since the function $\mathbf{1}_{\mathcal{U}_x}$ is the pointwise limit of an increasing sequence of such functions f, it follows that $n_\infty\{\mathcal{U}_x\}=0$. Furthermore, since S_i is contained in the union of a sequence of sets from $\{\mathcal{U}_x\colon x\in S_i\}$ it follows that $n_\infty\{S_i\}=0$. We conclude that

$$n_\infty\{\mathbf{X}\} = n_\infty\{\cup\, S_i\} + n_\infty\{\cap\, S_i^c\} = 0. \qquad \square$$

Observe that by Proposition 3.3, if (4) is w.s.c. then for each $x\in\mathbf{X}$

$$\frac{1}{N}\sum_{k=1}^{N}\mu_k \xrightarrow{\text{vaguely}} \alpha(x)\pi_x, \tag{43}$$

where, for $\alpha(x)\neq 0$, the invariant probability π_x is defined in (19). If $\alpha(x)=0$ then (43) still holds with $\pi_x = 0$.

The following result illustrates the sample path properties of w.s.c. Markovian systems.

PROPOSITION 3.4. *Suppose that conditions* A1–A4 *hold and that* Φ *is w.s.c. Then for any* $x \in \mathbf{X}$ *and* $f \in L^1(\mathbf{X}^{\mathbb{Z}_+}, \mathscr{B}(\mathbf{X}^{\mathbb{Z}_+}), P_{\pi_x})$,

$$P_x\left\{\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(\Phi_k, \Phi_{k+1}, \cdots) = f_\infty(\Phi_0, \Phi_1, \cdots)\right\} \geqq \alpha(x),$$

where $f_\infty \in L^1(\mathbf{X}^{\mathbb{Z}_+}, \mathscr{B}(\mathbf{X}^{\mathbb{Z}_+}), P_x)$ *is defined in* (20).

Proof. As before, let τ denote the first entrance time to H, and let $\mathscr{F}_\tau$ denote the sigma algebra of events before time τ.

By Proposition 2.3, for each $n \in I$, and $f \in L^1(\mathbf{X}^{\mathbb{Z}_+}, \mathscr{B}(\mathbf{X}^{\mathbb{Z}_+}), P_{\pi^n})$,

$$P_{\mu_0}\left\{\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(\Phi_k, \Phi_{k+1}, \cdots) = E_{\pi^n}[f(\Phi_0, \Phi_1, \cdots)]\right\} = 1$$

for any initial condition distribution μ_0 for wich $\mu_0\{H_n\} = 1$. From this and the strong Markov property it follows that for any $x \in \mathbf{X}$,

$$\begin{aligned}
&E_x\left[P_x\left\{\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(\Phi_k, \Phi_{k+1}, \cdots) = f_\infty(\Phi_0, \Phi_1, \cdots) \mid \mathscr{F}_\tau\right\} \mathbf{1}_{\tau<\infty}\right] \\
&= E_x\left[P_x\left\{\lim_{N\to\infty} \frac{1}{N} \sum_{k=\tau+1}^{\tau+N} f(\Phi_k, \Phi_{k+1}, \cdots) = f_\infty(\Phi_0, \Phi_1, \cdots) \mid \mathscr{F}_\tau\right\} \mathbf{1}_{\tau<\infty}\right] \\
&= E_x\left[P_{\Phi_\tau}\left\{\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(\Phi_k, \Phi_{k+1}, \cdots) = f_\infty(\Phi_0, \Phi_1, \cdots)\right\} \mathbf{1}_{\tau<\infty}\right] \\
&= P_x\{\tau < \infty\} = \alpha(x).
\end{aligned}$$

Hence

$$P_x\left\{\lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} f(\Phi_k, \Phi_{k+1}, \cdots) = f_\infty(\Phi_0, \Phi_1, \cdots)\right\} \geqq \alpha(x). \qquad \square$$

3.4. Proofs of Theorems 2.2–2.3. We begin with the proof of Theorem 2.2.

Proof of Theorem 2.2. We will proceed by establishing that (ii) implies (iii), and that (iii) implies (i). This is sufficient since it is obvious that (i) implies (ii), Propositions 3.3 and 3.4 show that (iii) implies (iv) and (v), and it is easy to see that (iv) and (v) each imply (ii).

The proof that the convergence in (v) is uniform on compact sets will be given in Proposition 3.5 below.

(ii)⇒(iii)

Let $\varepsilon > 0$, $x \in \mathbf{X}$ and $C \subset \mathbf{X}$ be a compact set for which

$$\liminf_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} P^k(x, C) > 1 - \varepsilon.$$

By Proposition 3.3 we have

$$\alpha(x) \geqq \alpha(x)\pi_x\{C\} = \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} P^k(x, C),$$

and since ε was chosen arbitrarily, this implies $\alpha(x) = 1$.

(iii)⇒(i)

Let $x \in \mathbf{X}$, $\varepsilon > 0$. Then by condition (iii) and Proposition 3.3

$$\mu_k = \sum_{i\in I} \alpha_i(x)\mu_k^i + n_k, \qquad \text{where } \alpha(x) = \sum_{i\in I} \alpha_i(x) = 1.$$

Choose $M \in \mathbb{Z}_+$ so large that $\sum_{i=0}^{M} \alpha_i > 1-\varepsilon$, and choose a compact set $C \subset \mathbf{X}$ such that

$$\gamma_k^i\{C\} > 1-\varepsilon, \quad \text{for } 0 \leqq i \leqq M, \quad \text{and} \quad 1 \leqq k \leqq \lambda^i.$$

Then by Proposition 3.3

$$\begin{aligned} \liminf_{k\to\infty} \mu_k\{C\} &\geqq \liminf_{k\to\infty} \sum_{i=1}^{M} \alpha_i \mu_k^i\{C\} + n_k\{C\} \\ &\geqq \sum_{i=1}^{M} \alpha_i \min_k \gamma_k^i\{C\} \qquad (44) \\ &\geqq (1-\varepsilon)^2. \end{aligned}$$

Hence (4) is bounded in probability. □

We now present a strengthening of Proposition 3.3 and show that the probabilities $\{\pi_x\}$ are continuous functions of $x \in \mathbf{X}$.

A set $A \in \mathcal{B}(\mathbf{X})$ will be called *uniform* if

$$\lim_{N\to\infty} \sup_{x\in A} \left\| \frac{1}{N} \sum_{k=1}^{N} P^k(x, \cdot) - \pi_x(\cdot) \right\|_{tv} = 0.$$

The following result is taken from [Cogburn, 1975].

LEMMA 3.8. *Suppose that conditions* A1–A4 *hold, and that* $\mathbf{\Phi}$ *is bounded in probability and w.s.c. Then for each* $i \in I$, *every compact subset of* S_i *is uniform.*

This will be used to obtain the following proposition.

PROPOSITION 3.5. *Suppose that conditions* A1–A4 *hold, and that* $\mathbf{\Phi}$ *is bounded in probability and w.s.c. Then every compact subset of* $\mathbf{X}$ *is uniform.*

Proof. Under the conditions of Proposition 3.5, $\alpha_i(x) = \Lambda(x, S_i)$, and since S_i is a Harris set and $\mathcal{V}_i \subset S_i$ has positive π_i-measure we have

$$\alpha_i(x) = \Lambda(x, \mathcal{V}_i). \qquad (45)$$

Let $C \subset \mathbf{X}$ be compact and $\varepsilon > 0$. Define the functions $h_N : \mathbf{X} \to \mathbb{R}_+$ for $N \in \mathbb{Z}_+$, and $x \in \mathbf{X}$ by

$$h_N(x) = P_x\left\{ \bigcup_{k=1}^{N} \bigcup_{i=0}^{N} \{\Phi_k \in \mathcal{V}_i\} \right\}.$$

By the Feller property, the function h_N is lower semicontinuous, and for each $x \in \mathbf{X}$, $h_N(x) \uparrow \alpha(x) = 1$ as $N \to \infty$ by (45). By Dini's theorem the convergence is uniform on compact subsets of $\mathbf{X}$.

Since the sets $\{S_i\}$ are absorbing, this implies that there exists an integer $N_0 \in \mathbb{Z}_+$ such that

$$P_x\left\{ \Phi_{N_0} \in \bigcup_{i=0}^{N_0} S_i \right\} \geqq h_{N_0}(x) > 1-\varepsilon$$

for all $x \in C$.

By the Feller property, the map $x \to P^{N_0}(x, \cdot)$ that takes $\mathbf{X}$ to $\mathcal{M}$, the space of probabilities on $\mathcal{B}(\mathbf{X})$, is continuous with respect to weak convergence on $\mathcal{M}$. Since the continuous image of a compact set is compact, the probabilities $\{P^{N_0}(x, \cdot) : x \in C\}$ are tight. Hence for some compact set $K \subset \mathbf{X}$,

$$P_x\left\{ \Phi_{N_0} \in \bigcup_{i=0}^{N_0} K \cap S_i \right\} > 1-\varepsilon \qquad (46)$$

for all $x \in C$.

For all $k\in\mathbb{Z}_+$, and $x\in\mathbf{X}$,

$$\left\|\frac{1}{N_0+k}\sum_{i=1}^{N_0+k} P^i(x,\cdot)-\pi_x(\cdot)\right\|_{tv}=\sup\left|\frac{1}{N_0+k}\sum_{i=1}^{N_0+k}\int P^i(x,dy)f(y)-\int\pi_x(dy)f(y)\right|,$$

where the supremum is taken over all measurable $f:\mathbf{X}\to[-1,1]$. Hence by (46) we have for all $x\in C$,

$$\begin{aligned}&\left\|\frac{1}{N_0+k}\sum_{i=1}^{N_0+k} P^i(x,\cdot)-\pi_x(\cdot)\right\|_{tv}\\
&\leqq\sup\int P^{N_0}(x,dz)\left|\frac{1}{k}\sum_{i=1}^{k}\int P^i(z,dy)f(y)-\int\pi_z(dy)f(y)\right|+\frac{N_0}{N_0+k}\\
&\leqq\sup\sum_{j=0}^{N_0}P^{N_0}(x,K\cap S_j)\sup_{z\in K\cap S_j}\left|\frac{1}{k}\sum_{i=1}^{k}\int P^i(z,dy)f(y)-\int\pi_z(dy)f(y)\right|\\
&\qquad+\frac{N_0}{N_0+k}+\varepsilon\\
&\leqq\sup\sum_{j=0}^{N_0}\sup_{z\in K\cap S_j}\left|\frac{1}{k}\sum_{i=1}^{k}\int P^i(z,dy)f(y)-\int\pi_z(dy)f(y)\right|+\frac{N_0}{N_0+k}+\varepsilon\\
&\leqq\sum_{j=0}^{N_0}\sup_{z\in K\cap S_j}\left\|\frac{1}{k}\sum_{i=1}^{k}P^i(z,\cdot)-\pi_z(\cdot)\right\|_{tv}+\frac{N_0}{N_0+k}+\varepsilon.\end{aligned}$$

Applying Lemma 3.8 to the inequality above shows that

$$\limsup_{N\to\infty}\sup_{x\in C}\left\|\frac{1}{N}\sum_{i=1}^{N}P^i(x,\cdot)-\pi_x(\cdot)\right\|_{tv}\leqq\varepsilon.$$

Since ε is arbitrary, this proves the proposition. □

The following result shows that when $\mathbf{\Phi}$ satisfies the conditions of Proposition 3.5, the map $x\to\pi_x$ that takes $\mathbf{X}$ to $\mathcal{M}$ is continuous in total variation norm. In particular, the kernel $\Pi=\pi_{(\cdot)}(\cdot)$ has the strong Feller property.

COROLLARY 3.1. *Suppose that conditions* A1–A4 *hold. If* $\mathbf{\Phi}$ *is bounded in probability and w.s.c. then for each* $i\in I$, *the function* $\alpha_i(x)$ *is a continuous function of* $x\in\mathbf{X}$.

Proof. Fix $i\in I$, and let f_i be a continuous function for which $0\leqq f_i\leqq 1$, and

$$f_i(x)=\begin{cases}1 & \text{if } x\in S_i;\\ 0 & \text{if } x\in S_j, \qquad j\neq i.\end{cases}$$

It follows from Theorem 2.2 and the definition of π_x that

$$\alpha_i(x)=\int f_i\,d\pi_x=\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N}\int P^k(x,dy)f_i(y),$$

where the convergence is uniform for x in compact subsets of $\mathbf{X}$. The corollary follows from these facts since for each $N\in\mathbb{Z}_+$, the function

$$\frac{1}{N}\sum_{k=1}^{N}\int P^k(\cdot,dy)f_i(y)$$

is continuous. □

Proof of Theorem 2.3. Suppose that $\mathbf{\Phi}$ is uniformly bounded in probability, and let $C\subset\mathbf{X}$ be a compact set for which

$$\liminf_{k\to\infty} P^k(x, C)\geqq\tfrac{1}{2}\quad\text{for every } x\in\mathbf{X}.$$

By Proposition 3.3, for each $x\in\mathbf{X}$,

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} P^k(x, C)=\pi_x\{C\},$$

and hence $\pi^i\{C\}\geqq\frac{1}{2}$ for each $i\in I$. However by Lemma 3.4

$$\pi^i\{C\}=\pi^i\{C\cap S_i\}=0$$

for all but a finite number of $i\in I$ and this shows that I is in fact finite.

Conversely, if I is finite and $\mathbf{\Phi}$ is bounded in probability then Proposition 3.3 shows that the distributions of $\mathbf{\Phi}$ converge cyclically. Since in this case the collection of probabilities $\{\pi_x\colon x\in\mathbf{X}\}$ is tight, it follows that $\mathbf{\Phi}$ is uniformly bounded in probability.

To prove that the distributions of $\mathbf{\Phi}$ converge uniformly on compact sets, use the proof of Proposition 3.5. □

Proof of Theorem 2.4. The main idea of the proof of Theorem 2.4 is that condition GA or A5 ensures that the support of every invariant probability contains the origin. Under the weak stochastic controllability hypothesis, the supports of distinct the invariant probabilities $\{\pi^i\}$ are disjoint, and hence Harris recurrence is an immediate consequence. The idea of replacing a random disturbance by a deterministic sequence lying in its support has previously been used to determine the support of a diffusion process (see [Stroock and Varadhan, 1972] and [Kunita, 1976]).

To prove (i) we observe that the conditions of Theorem 2.2 are satisfied and hence by Corollary 2.3 it is enough to show that there is exactly one invariant probability.

Let π be one of the invariant probabilities $\{\pi^i\}$, and $\mathcal{W}$ any open set that contains the origin. By Lemma 2.1, the sets

$$D_n=\{x\in\mathbf{X}\colon P^n(x,\mathcal{W})>0\}\qquad n\in\mathbb{Z}_+$$

cover $\mathbf{X}$, and hence for some $n\in\mathbb{Z}_+$, $\pi\{D_n\}>0$. This implies that $\pi\{\mathcal{W}\}\geqq\int_{D_n} P^n(y,\mathcal{W})\pi(dy)>0$, and since $\mathcal{W}$ is an arbitrary open set containing the origin, it follows that $0\in\operatorname{supp}\pi$. By Lemma 3.3 the supports of π^i and π^j are disjoint for $i\neq j$, and this proves (i) of the proposition.

To prove (ii) suppose that a cycle $\mathbf{E}$ exists with period λ, let π denote the unique invariant probability for $\mathbf{\Phi}$, and let $\mathcal{U}$ and $\mathcal{V}$ be open sets that satisfy condition (iii) of Proposition 2.1.

Let $d_k\triangleq\lambda\mathbf{1}_{E_k}\pi$. Since the sets in the cycle $\mathbf{E}$ are disjoint, d_k and d_j are mutually singular for $k\neq j\,(\operatorname{mod}\lambda)$. However by Lemma 2.1 we have $d_k\{\mathcal{U}\}=\int P^k(x,\mathcal{U})d_0(dx)>0$ for all k sufficiently large, and this implies that $d_k>\mathbf{1}_{\mathcal{V}}\mu^{\text{Leb}}$ for all large k. Since $\{d_k\colon 1\leqq k\leqq\lambda\}$ are mutually singular, this implies that $\lambda=1$. □

4. Conclusions. In this paper we have presented a stability theory for nonlinear stochastic systems operating under feedback. Among the consequences of this theory are the following:

(i) If a Markovian system of the form (4) is bounded in probability and w.s.c.

then the limits

$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} f(\Phi_k)=f_\infty \quad \text{a.s. } [P_x],$$
(47)
$$\lim_{N\to\infty}\frac{1}{N}\sum_{k=1}^{N} P_x\{\Phi_k\in A\}=\pi_x\{A\}$$

exist for *every* initial condition $x\in\mathbf{X}$, every $A\in\mathscr{B}(\mathbf{X})$, and for a very large class of functions f that includes all positive Borel measurable functions. This result is sharp because in the case where $F(\cdot,\cdot)$ is linear but not controllable there may exist Borel sets A and bounded Borel functions f for which (47) does not hold.

(ii) If (4) is uniformly bounded in probability and w.s.c. then the probabilities $\{P^k\colon k\in\mathbb{Z}_+\}$ governing the state process converge cyclically for *any* initial condition. For example, this implies that there exists a $\lambda\in\mathbb{Z}_+$ such that

$$\lim_{k\to\infty} P\{\Phi_{k\lambda+i}\in A\}$$

exists for every Borel measurable set $A\subset\mathbf{X}$ and every $i\in\mathbb{Z}_+$.

(iii) Under a mild stability condition on the undriven system, periodicity is ruled out and a unique invariant probability π may be shown to exist. Hence, for example, the limit above holds with $\lambda=1$.

These results have significant implications in stochastic system theory and, in particular, to the control of stochastic systems that vary randomly in time.

To illustrate what can go wrong with stochastic controllability, consider a single input single output ARMAX system under the mean-square optimal control law

$$B(z)u(z)=[A(z)-C(z)]y(z),$$

and define

$$\Phi_{k+1}=(y_k,\cdots,y_{k-n_1},u_{k-d},\cdots,u_{k-n_2},w_k,\cdots,w_{k-n_3})^{\mathrm{T}}.$$

Suppose that the zeros of the polynomials B and C lie outside the unit disc in $\mathbb{C}$. It is easily shown that the closed loop system is bounded in probability in this case and that

$$\lim_{k\to\infty}(y_k-w_k)=0 \quad \text{a.s. } [P_{\Phi_0}] \tag{48}$$

for all initial conditions Φ_0.

By (48) we have $y_0=w_0$ almost surely $[P_\pi]$ for the (unique) invariant probability π, and hence π is supported on a hyperplane in $\mathbb{R}^n$. Consequently, this system is *not* w.s.c. because π is singular with respect to Lebesgue measure. Similarly, the stochastic gradient algorithm of [Goodwin, Ramadge, and Caines, 1981] does not give rise to a w.s.c. system because the gain r_k^{-1} converges to zero almost surely.

However one active research area in stochastic control theory today is the adaptive control of time varying stochastic systems (see for example [Chen and Caines, 1985]). It is in this area that the ideas introduced in this paper will be very useful. For example, the Markovian system of [Meyn and Caines, 1987] is l.s.c., and this fact was crucial in establishing many of the results in that paper. Furthermore, an ARMAX system controlled by a gradient type algorithm gives rise to a w.s.c. Markovian system under mild conditions [Meyn and Guo, 1990].

There is much room for future research. For example, the asymptotic behavior of Markovian systems that are not w.s.c. is not well understood. We have already noted that in the linear case, $\Phi_{k+1}=A\Phi_k+Bw_{k+1}$, there are two extreme cases when A is asymptotically stable and the disturbance $\mathbf{w}$ satisfies condition A4. If (A, B) is controllable then $\mathbf{\Phi}$ is positive Harris recurrent and hence the law of large number holds for every positive Borel function on the state space and every initial condition distribution. If (A, B) is not controllable then this is not the case except in trivial situations (when A and B are both zero, for instance). However the law of large numbers does hold for continuous functions with arbitrary initial distributions.

There is evidence to suspect that this result may be generalized to arbitrary Feller Markov chains that are bounded in probability and possess exactly one invariant probability: By Proposition 2.2, for any such Markov chain and any bounded and continuous function f on the state space $\mathbf{X}$ we have

$$\lim_{N\to\infty} E_x\left[\frac{1}{N}\sum_{k=1}^{N} f(\Phi_k)\right]=\int f\,d\pi.$$

This suggests that it may be possible to remove the expectation operator "E_x" to establish the law of large numbers for the class of bounded and continuous functions.

One approach to the problem is to search for stronger stability assumptions on $\mathbf{\Phi}$ that allow an approximation with a w.s.c. Markov chain $\mathbf{\Phi}^{\varepsilon}$.

Another possible approach is to use some nonlinear version of the controllable/uncontrollable decomposition theorem used in linear system theory. By applying the results of this paper to the controllable part of $\mathbf{\Phi}$, it may be possible to achieve the desired ergodic properties.

For example, a result of [Jakubczyk and Sontag, 1989] shows that if a fixed point exists (a state $x\in\mathbf{X}$ with the property that $F(x, w)=x$ for some $w\in \operatorname{supp}\mu_w$) then under general conditions an invariant submanifold $x\in\mathbf{Y}\subset\mathbf{X}$ exists on which the restricted system is w.s.c. Applying the results of this paper, it is easily shown that if such a system is bounded in probability then an aperiodic Harris set containing x exists. This allows an analysis of the asymptotic behavior of $\mathbf{\Phi}$, at least in the case when $\Phi_0=x$. Some ideas along these lines, and extensions to arbitrary initial conditions were begun in [Meyn, 1989].

REFERENCES

A. Aloneftis (1987), *Stochastic Adaptive Control Results and Simulations*, Lecture Notes in Control and Information Sciences, Springer-Verlag, New York.

A. Arapostathis and S. I. Marcus (1990), *Analysis of an identification algorithm arising in the adaptive estimation of Markov chains*, Mathematics of Control, Signals, and Systems, 3, pp. 1-29.

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.

P. E. Caines (1988), *Linear Stochastic Systems*, John Wiley, New York.

R. Cogburn (1975), *A uniform theory for sums of Markov chain transition probabilities*, Ann. Probab., 3, pp. 191-214.

H. F. Chen and P. E. Caines (1985), *On the adaptive control of a class of systems with random parameters and disturbances*, Automatica, 21, pp. 737-741.

J. L. Doob (1953), *Stochastic Processes*, John Wiley, New York.

Feller (1971), *An Introduction to Probability Theory and its Applications, Volume* 2, John Wiley, New York.

E. Fernandez-Gaucherand, A. Arapostathis, and S. I. Marcus (1988), *On the adaptive control of a partially observable Markov decision process*, Proc. 27th IEEE Conference on Decision and Control, Austin, TX, December 7-9, pp. 1204-1210.

S. R. Foguel (1969), *Positive operators on $C(X)$*, Proc. AMS, 22, pp. 295-297.

——— (1973), *The ergodic theory of positive operators on continuous functions*, Ann. Scuola Norm. Sup. Pisa, 27, pp. 19-51.

G. C. GOODWIN, P. J. RAMADGE, AND P. E. CAINES (1981) *Discrete time stochastic adaptive control*, SIAM J. Control Optim., 19, p. 829, Corrigendum, 20, 1982, p. 893.

R. Z. HAS'MINSKIĬ (1980), *Stochastic stability of differential equations*, Sitjhoff and Noordhoff Alphen an den Rijn, the Netherlands; Rockville, Maryland.

L. GUO AND S. P. MEYN (1989), *Adaptive control for time-varying systems: a combination of martingale and Markov chain techniques*, International J. of Adaptive Control and Signal Processing, 3, pp. 1–14.

K. ICHIHARA AND H. KUNITA (1974), *A classification of the second order degenerate elliptic operators and its probabilistic characterization, Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 30, pp. 235–254.

B. JAKUBCZYK AND E. D. SONTAG (1990), *Controllability of nonlinear discrete time systems: a Lie-algebraic approach*, SIAM J. Control Optim., 28, pp. 1–33.

KLIEMANN (1987), *Recurrence and invariant measures for degenerate diffusions*, Ann. Probab., 15, pp. 690–707.

P. R. KUMAR (1985), *A survey of some results in stochastic adaptive control*, SIAM J. Control Optim., 23, pp. 329–380.

H. KUNITA (1976), *Suppports of diffusion processes and controllability problems*, Proc. of Intern. Symp. SDE, Kyoto, Japan, pp. 163–185.

H. J. KUSHNER AND A. SHWARTZ (1984), *An invariant measure approach to the convergence of stochastic approximations with state dependent noise*, SIAM J. Control Optim., 22, pp. 13–27.

H. J. KUSHNER (1972), *Stochastic stability*, in Lecture Notes in Mathematics, vol. 294, Springer-Verlag, New York.

——— (1967), *Stochastic stability and control*, Academic Press, New York.

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

S. P. MEYN AND R. L. TWEEDIE (1990a), *Criteria for stability of Markovian processes* I: *discrete time chains*, J. Appl. Prob., to appear.

——— (1990b), *Criteria for stability of Markovian processes* II: *continuous time processes*, submitted.

S. P. MEYN AND L. GUO (1990), *Stability, convergence, and performance of an adaptive control algorithm applied to a randomly varying system*, to appear IEEE Trans. Automat. Control.

S. P. MEYN AND P. E. CAINES (1987), *A new approach to stochastic adaptive control*, IEEE Trans. Automat. Control, AC-32, pp. 220–226.

E. NUMMELIN (1984), *General Irreducible Markov Chains and Non-Negative Operators*, Cambridge University Press, Cambridge, UK.

S. OREY (1971), *Limit Theorems for Markov Chain Transition Probabilities*, Van Nostrand Reinhold Mathematical Studies 34, Van Nostrand, London.

S. H. SAPERSTONE (1981), *Semidynamical Systems in Infinite Dimensional Spaces*, Springer-Verlag, Berlin, New York.

E. D. SONTAG (1983), *A Lyapunov-like characterization of asymptotic controllability*, SIAM J. Control Optim., 21, pp. 462–471.

D. W. STROOCK AND S. R. VARADHAN (1972), *On the support of diffusion processes with applications to the strong maximum principle*, Proc. 6th Berkeley Sympos. Math. Statist. Prob., pp. 333–368.

P. TUOMINEN AND R. L. TWEEDIE (1979), *Markov chains with continuous components*, Proc. London Math. Society, Series 3, Vol. 38, pp. 89–114.

W. M. WONHAM (1966), *A Liapunov criteria for weak stochastic stability*, J. Differential Equations, 2, pp. 195–207.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 3, pp. 562-593, May 1991

004

EXTENDED ZEROS AND MODEL MATCHING*

MICHAEL K. SAIN†, BOSTWICK F. WYMAN‡, AND JOSEPH L. PECZKOWSKI§

Abstract. The model matching equation $T(z)=P(z)M(z)$ induces constraints upon the multivariate zero structures of $P(z)$ and $M(z)$; the nature of the constraint is best explained by extending the usual notion of zero. In particular, the extended Γ-zero module of $P(z)$ must contain as a submodule the module Z_Γ of matching Γ-zeros, which depends only upon $T(z)$ and $M(z)$; and the extended Ω-zero module of $M(z)$ must contain as a factor module the module Z_Ω of matching Ω-zeros, which depends only upon $T(z)$ and $P(z)$. Essential solutions, in which the constraint is by module isomorphism, are possible if and only if the nullity of $P(z)$ does not exceed the nullity of $T(z)$, on the one hand, or the co-nullity of $M(z)$ does not exceed the co-nullity of $T(z)$, on the other. Both the matching zero modules and their finitely generated, torsion parts—which have state-space interpretation—can be given concrete, intuitive interpretation in terms of short exact sequences, though the former is less involved than the latter. Moreover, in the case of the latter, a natural notion of essential solution is not available, in marked contrast to the situation for poles.

Key words. zeros, zero modules, extended zeros, model matching

AMS(MOS) subject classifications. 93B25, 13C10

1. Introduction.

We begin by considering the linear transfer function equation

$$T(z)=P(z)M(z), \tag{1.1}$$

over the field $k(z)$ of rational functions in z having coefficients in a base field k. If $R(z)$, $U(z)$, and $Y(z)$ are the vector spaces, then (1.1) may be seen either in terms of the commutative diagram of Fig. 1.1 or that of Fig. 1.2, depending upon whether $T(z)$ and $M(z)$ are given and $P(z)$ is sought, or $T(z)$ and $P(z)$ are given and $M(z)$ is sought.

The problems of Figs. 1.1 and 1.2 arise naturally in various applications. For example, in the theory of codes for reliable communication, $P(z)$ has to do with the decoding process, while $M(z)$ is associated with code design (see [1] for additional discussion). Again, in the theory of feedback control for solutions to servomechanism problems, $P(z)$ may represent the plant and $M(z)$ the open-loop equivalent of precompensation and feedback; or $M(z)$ may represent the plant and $P(z)$ the open-loop equivalent of post-compensation and feedback (for illustrations, see [2]).

$$\begin{tikzcd} & R(z) \arrow[dl, "M(z)"'] \arrow[dr, "T(z)"] & \\ U(z) \arrow[rr, dashed, "P(z)"'] & & Y(z) \end{tikzcd}$$

FIG. 1.1

$$\begin{tikzcd} & R(z) \arrow[dl, dashed, "M(z)"'] \arrow[dr, "T(z)"] & \\ U(z) \arrow[rr, "P(z)"'] & & Y(z) \end{tikzcd}$$

FIG. 1.2

* Received by the editors August 8, 1988; accepted for publication (in revised form) June 26, 1990.

† Department of Electrical and Computer Engineering, University of Notre Dame, Notre Dame, Indiana 46556. The work of this author was supported by The Ohio State University's Distinguished Visiting Professorship, by the University of Notre Dame's Frank M. Freimann Chair in Electrical Engineering, and by National Science Foundation grant ECS-84-05714.

‡ Department of Mathematics, The Ohio State University, Columbus, Ohio 43210.

§ Energy Controls Division, Allied-Bendix Aerospace, South Bend, Indiana 46620.

In feedback theory, (1.1) has come to be known as the model matching equation. Accordingly, Figs. 1.1 and 1.2 represent two versions of the model matching problem. The model matching problem has been a central issue of abstract control theory at least since the seminal work [3] of Wang and Davison in 1973. For a sample of technically and historically important papers in this area, see also [4]-[6]. Unfortunately, the diversity of language and points of view in this area makes a useful survey and comparison of results very difficult, and we cannot undertake a thorough survey in this already long and technical paper. Nevertheless, later in this Introduction, we discuss briefly the differences in outlook between the present paper and some previous work.

If nothing is required of the unknown mappings $P(z)$ in Fig. 1.1 and $M(z)$ in Fig. 1.2 beyond their existence as $k(z)$-linear transformations, then almost trivial necessary and sufficient conditions are well known. In Fig. 1.1, given $T(z)$ and $M(z)$, then there exists a $k(z)$-linear map $P(z)$ completing the diagram if and only if the nullspace $\ker M(z)$ is a subspace of $\ker T(z)$. In Fig. 1.2, there exists a $k(z)$-linear map $M(z)$ completing the diagram if and only if the image $T(z)R(z)$ is a subspace of $P(z)U(z)$. If these basic conditions are satisfied, then in general the solutions are not unique, and the interesting problems center around constraints describing the set of possible solutions. In this Introduction we discuss the constraints which arise in Fig. 1.1, although both diagrams are discussed in the paper.

Given $T(z)$ and $M(z)$, assuming the kernel condition above, what mappings $P(z)$ can occur as solutions? We approach this problem by trying to find out what poles and what zeros are required in any solution mapping $P(z)$, and we call these minima the fixed poles and the fixed zeros.

The problem of fixed poles in this context was discussed by Conte, Perdon, and Wyman [7], building on earlier work in [8] and [9]. For a solution $P(z)$, we denote the pole module of $P(z)$ by $P(P(z))$. By this we mean that $P(P(z))$ is the state space of the minimal realization of $P(z)$ considered as a module over the polynomial ring $k[z]$ in the usual way. In [7] a new module

$$P(T, M) = M^{-1}U[z]/\{M^{-1}U[z] \cap T^{-1}Y[z]\} \tag{1.2}$$

was introduced, and the first basic result there established that $P(T, M)$ must appear as a submodule of $P(P(z))$ for any solution $P(z)$. Now the module $P(T, M)$ is a finite-dimensional vector space with a dynamics matrix arising from the module action, and the eigenvalues of that matrix give the numerical values of the poles, so we know numerically what poles must exist in $P(z)$. A module inclusion result gives substantially deeper additional information about multiplicities and Jordan structure of the space of poles. Furthermore, "essential solutions" satisfying exactly $P(P(z)) = P(T, M)$ can be constructed; and the module $P(T, M)$ can be given a precise and attractive description in terms of the pole and zero modules of $T(z)$ and $M(z)$. Finally, this method is not limited solely to the study of finite poles and zeros, but by replacing the polynomial ring by the ring of proper rational functions or suitable more complicated rings, conclusions can be drawn about causal or causal and stable solutions by considering poles and zeros at infinity. One of the chief dividends of the present algebraic approach is its ability to attack varying problems (such as causality and various kinds of stability) by simply changing the rings involved. The polynomial ring can be replaced by the ring of proper or proper and stable rational functions. Alternatively, a solution to the model matching problem is causal if it has no poles at infinity, and it is causal and stable if it has no poles at infinity or in the designated unstable region of the complex plane. For further discussion of this philosophy, see § 9 herein.

The present paper considers the problem of "fixed zeros" of the desired solution mappings $P(z)$. The goal is to construct a module, say $Z(T, M)$, which would appear as a submodule of the transmission zero module of any $P(z)$. However, the theory of zeros is much more complicated than the theory of poles; and the results obtained must be expressed in module language, having in general no adequate numerical version. The major source of the difficulties is in fact the nullspace of $T(z)$, which creates an infinite-dimensional space of "generic zeros" which can nevertheless affect the lumped or finite-dimensional zero spaces of $P(z)$.

At this point it will be helpful to discuss the differences in outlook between the present work and that part of the model matching tradition represented by the important paper [10] of Emre. Emre considers the problem $T(z)=P(z)M(z)$ as in Fig. 1.1 above, where $M(z)$ and $T(z)$ are given and $P(z)$ is sought. This problem is called the exact model matching problem (EMMP), and the "generalized EMMP" adds a causality requirement. The solution in [10] is quite explicit, supplying algorithms and parameterized families of solutions. On the other hand, [10] is primarily concerned with the pole structure of the solutions, generalizing [8] in a way different from the results in [7]. Work on zeros in this spirit can be found in [11] which, however, does not discuss model matching explicitly (see also the classic [12]). We are grateful to the referees for references to the closely related dynamic cover problem. See [10, p. 157] and Antoulas' paper [13]. We should point out that there is a close relationship between model matching, zeros, and the geometric control theory. This is evident, not only in [7]-[9], but also in [14], which gives an explicit map between the zero module and spaces of the form V^*/R^*. Finally, we mention that the zeros of a composition $P(z)M(z)$ of two given $k[z]$-linear maps received earlier attention in [15] and [16]. Neither of these, however, addresses model matching.

We hope that the present work makes a new contribution to the diversity of approaches to the model matching problem. The technical results on fixed zeros are new, but perhaps difficult to relate to earlier work. We hope that this paper motivates some reader to prepare a comprehensive survey of all these competing points of view.

Finally, we supply a "roadmap" to the results in this paper. Section 2 is a review of the basic ideas surrounding the pole module and the finite-dimensional zero module of a transfer function. In § 3 these ideas are generalized to ideas involving certain (possibly infinite-dimensional) extended zero modules which arise inevitably in the study of transfer functions which fail to be injective and surjective. In § 4, the heart of the paper, new modules describing the "matching zero structures" associated to Figs. 1.1 and 1.2 are introduced. For example, the relevant matching zero module for Fig. 1.1 (defined in (4.5) below) is

$$Z_\Gamma = T^{-1}Y[z]/\{T^{-1}Y[z]\cap M^{-1}U[z]\}. \tag{1.3}$$

Although this module appears very similar to the $P(T, M)$ defined above for the fixed pole problem, it is in fact much more complicated. If $T(z)$ is not one-to-one, the module Z_Γ may contain an infinite-dimensional part (which is actually a divisible module over $k[z]$) coming from the nullspace of $T(z)$. The results obtained in this paper indicate that this phenomenon cannot be ignored. The central result of § 4 states in this case (Theorem 3(2)) that Z_Γ appears as a submodule of the (possibly infinite-dimensional) extended zero module $Z_\Gamma(P(z))$, so that we must consider Z_Γ as the most natural module of "fixed zeros" for the problem. Section 5 contains an extensive analysis of Z_Γ in terms of more familiar and intuitive concepts. Section 6 investigates the problem of "essential solutions," seeking solutions $P(z)$ such that Z_Γ exactly equals $Z_\Gamma(P(z))$, concluding that they do not in general exist. Theorem 6 gives necessary and sufficient conditions for the existence of essential solutions. Sections 7 and 8 discuss the complications which arise when we try to prove a corresponding theorem for the finite-dimensional matching zero spaces and the finite-dimensional transmission zero module. Although elaborate and detailed results are obtained here, it is hard to imagine discovering and proving them directly in terms of matrix theory and finite-dimensional lumped zero theory. We are convinced more than ever that the module-theoretic approach to zeros, including now the infinite-dimensional part of the theory, gives the "correct" approach to problems of the type considered here.

Although this Introduction has considered only the form of the problem arising from Fig. 1.1 and leading to divisible modules of extended zeros, the problem formulation of Fig. 1.2 is also examined in the paper. This version requires infinite-dimensional free modules of extended zeros and a matching zero module with possibly a free part. We remark that it is in principle possible to handle Fig. 1.2 by a sophisticated dualization of Fig. 1.1. Our experience in looking at this approach leads us to conclude that it increases technicality without saving space in a comparable way. Moreover, it reduces the possibility of giving intuitive interpretations to the building blocks of these modules. We have therefore chosen a parallel treatment.

2. Finitely generated, torsion poles and zeros. We set the following conventions. For a field k, let $k[z]$ be the ring of polynomials in z with coefficients in k, and let $k(z)$ be the induced quotient field. Recall that a "transfer function" in $k(z)$ is an equivalence class, and note that the customary representative of a class is a pair $(n(z), d(z))$ of polynomials in $k[z]$, with $d(z)$ nonzero and the pair relatively prime. In applications, the ratio $n(0)/d(0)$ is also of interest, when it is defined, and is designated as the "gain" K of the associated class. More generally, we can represent a transfer function in the manner $(Kz^{\alpha}n(z), d(z))$ or $(Kn(z), z^{\beta}d(z))$, where there is the additional condition that $n(0)=d(0)=1$. For the purposes of the investigation in this paper, any unit in $k[z]$ will serve just as well as any other. With these understandings, we refer to elements in $k(z)$ as transfer functions.

For a discussion of poles and zeros in the multivariable sense, it is necessary to have in hand a concrete representation of polynomials and transfer functions in a vector sense. Intuitively, we do this by "multiplying" the scalar version of the quantity in question onto the vector of interest. More precisely, let R, U, and Y be finite-dimensional vector spaces over k; and note that $k[z]$ and $k(z)$ are also k-vector spaces. Then we employ the k-bilinear tensor product to form $k[z]$-modules

$$(2.1) \qquad R[z]=k[z]\otimes_k R, \quad U[z]=k[z]\otimes_k U, \quad Y[z]=k[z]\otimes_k Y,$$

and $k(z)$-vector spaces

$$(2.2) \qquad R(z)=k(z)\otimes_k R, \quad U(z)=k(z)\otimes_k U, \quad Y(z)=k(z)\otimes_k Y.$$

If $M(z): R(z)\rightarrow U(z)$ is a $k(z)$-linear map, we wish to describe its poles and zeros in a precise, technical sense. For the elementary class represented by $(n(z), d(z))$, this is done easily and intuitively with the $k[z]$-modules $k[z]/d(z)k[z]$ and $k[z]/n(z)k[z]$. For $M(z)$, greater elaboration is needed.

The pole module $P(M(z))$ associated with a $k(z)$-linear map $M(z)$ is defined to be the $k[z]$-factor module given by

$$(2.3) \qquad P(M(z))=R[z]/\{R[z]\cap M^{-1}U[z]\},$$

where M^{-1} is the inverse image function of $M(z)$, defined on $U[z]$ by

$$(2.4) \qquad M^{-1}U[z]=\{r(z): r(z)\in R(z) \text{ and } M(z)r(z)\in U[z]\}.$$

As a $k[z]$-module, $P(M(z))$ is finitely generated, because $R[z]$ is finitely generated; it is a torsion module because every element in $R[z]$, even if not in $M^{-1}U[z]$, can be made so by scaling with a suitable polynomial $p(z)$ in $k[z]$.

Finitely generated, torsion modules over $k[z]$, such as the pole module $P(M(z))$, can be understood as traditional state spaces of finite dimension over the field k. Considered as a k-vector space, $P(M(z))=X_P$, the usual state space of a minimal realization of $M(z)$; and the scalar module action $z:P(M(z))\rightarrow P(M(z))$ defines a k-linear dynamics map $A_P:X_P\rightarrow X_P$ in the realization. On the system level, it may be observed from the definition (2.3) that the input signals of interest are polynomial vectors in $R[z]$, but not those which produce output vectors which lie entirely in $U[z]$. In terms of classical realization theory [17], entries in $R[z]$ may be understood as finite sequences of inputs starting in the past and ending at the present. An output contained entirely in $U[z]$ corresponds to a sequence all of whose future values are zero. From an engineering point of view, we may think of exciting the system with inputs which do not have poles, but which lead to outputs which do have poles.

In a quite realistic sense, we may regard the pole module $P(M(z))$ as a generalization of the "denominator polynomial" $d(z)$. The use of a module such as (2.3) permits us to carry information about the poles of $M(z)$ in a capsule form. The module describes, simultaneously, the dimension of the state space and the spectral character of its associated dynamics.

It is sometimes convenient to make use of isomorphic forms for the pole module (2.3). For example, we can write

$$P(M(z))\approx MR[z]/\{U[z]\cap MR[z]\}, \tag{2.5}$$

as a $k[z]$-module. To develop such an isomorphism, we can begin with the commutative diagram of Fig. 2.1, in which rows are natural inclusions and columns are restrictions of $M(z)$. If $r(z)$ represents an element in (2.3), then column two of Fig. 2.1 induces the desired isomorphism by the action

$$r(z)\mapsto M(z)r(z) \bmod MR[z]\cap U[z]. \tag{2.6}$$

The pole module of $M(z)$ maintains its basic character, being finitely generated and torsion, even when the kernel, $\ker M(z)$, of $M(z)$, or the cokernel, $\operatorname{coker} M(z)$, of $M(z)$ is nonzero. It turns out that this is not the situation for $k[z]$-zeros of $M(z)$. However, there is a fundamental module containing the zeros of $M(z)$ which are finitely generated and torsion. These zeros, of course, will be the most familiar because they resemble poles.

The finitely generated, torsion zero module $Z(M(z))$ associated with a $k(z)$-linear map $M(z)$ was defined by Wyman and Sain [8] in 1981 to be the $k[z]$-factor module

$$Z(M(z))=\{M^{-1}U[z]+R[z]\}/\{\ker M(z)+R[z]\}. \tag{2.7}$$

$$\begin{tikzcd}
R[z]\cap M^{-1}U[z] \arrow[r] \arrow[d, "M(z)"'] & R[z] \arrow[d, "M(z)"] \\
MR[z]\cap U[z] \arrow[r] & MR[z]
\end{tikzcd}$$

FIG. 2.1

From the fact that $R[z]$ and $U[z]$ are finitely generated, together with the fact that elements are equivalent modulo $\ker M(z)$, we have that (2.7) is finitely generated. It is torsion because any element in $M^{-1}U[z]$ can be scaled by a polynomial $p(z)$ in $k[z]$ so that the result lies in $R[z]$. From the definition, it is clear that the module $Z(M(z))$ has to do with inputs in $R(z)$ which lead to outputs in $U[z]$, but not polynomial vector inputs in $R[z]$ and not inputs in $\ker M(z)$. Intuitively, then, we are talking about system excitations which contain poles, but which produce nonzero responses without poles. In the sequel, when we use the term zero module, we will have in mind the finitely generated, torsion module (2.7), or one of its isomorphic forms. Two such forms are given by

$$Z(M(z)) \approx M^{-1}U[z]/\{M^{-1}U[z] \cap (\ker M(z) + R[z])\} \tag{2.8}$$

$$\approx \{U[z] \cap MR(z)\}/\{U[z] \cap MR[z]\}; \tag{2.9}$$

and explicit isomorphisms corresponding to (2.8) and (2.9) can be determined from the commutative diagrams of Figs. 2.2 and 2.3, respectively.

In a manner analogous to poles, the zero module permits a state-space interpretation. As a k-vector space, $Z(M(z)) = X_Z$, which is of finite dimension. Also, the scalar module action $z : Z(M(z)) \to Z(M(z))$ defines a k-linear map $A_Z : X_Z \to X_Z$. The space X_Z is not the usual state space; and the map A_Z is not the usual dynamics map. In order to distinguish the cases, we can think of X_P as a space of pole-states and of X_Z as a space of zero-states. Then A_P becomes the pole-state dynamics map, while A_Z is the zero-state dynamics map. It should be noted that the terminology "zero-state" is not in reference to any initial conditions. Rather, we are speaking of two distinct state spaces which are associated with the same $k(z)$-linear map. One of these spaces is for poles, while the other is for zeros.

The module $Z(M(z))$ is a generalization of the elementary notion of "numerator polynomial" $n(z)$, even as the module $P(M(z))$ generalizes $d(z)$. The algebra of these modules is somewhat more intricate than that of the ring $k[z]$, however, and involves commutative diagrams and exact sequences. Moreover, for the study of the problem in this paper, other types of $k[z]$-zeros have an important role to play. Some of these zeros are of torsion type, but not finitely generated. Others are finitely generated, but

$$\begin{tikzcd}
M^{-1}U[z] \cap (\ker M(z) + R[z]) \arrow[r] \arrow[d] & M^{-1}U[z] \arrow[d] \\
\ker M(z) + R[z] \arrow[r] & M^{-1}U[z] + R[z]
\end{tikzcd}$$

FIG. 2.2

$$\begin{tikzcd}
M^{-1}U[z] \cap (\ker M(z) + R[z]) \arrow[r] \arrow[d, "M(z)"'] & M^{-1}U[z] \arrow[d, "M(z)"] \\
U[z] \cap MR[z] \arrow[r] & U[z] \cap MR(z)
\end{tikzcd}$$

FIG. 2.3

are not torsion. We have, therefore, to extend the notion of zero; and this is the content of § 3.

3. Extended zeros: divisible and free. In § 2, we discussed $k[z]$-zeros of a $k(z)$-linear map $M(z): R(z) \to U(z)$. These zeros were finitely generated and torsion; and, accordingly, they possess a natural interpretation of state-space type. Such zeros, which have received almost all the attention in the systems literature, are most familiar because of their resemblance to poles. For systems having one input and one output, this concept of zero is entirely adequate. When multiple inputs and multiple outputs are considered, however, zeros appear which are not of pole-type. Some of these are torsion zeros, but not finitely generated; in the context of realization theory, these may be classified as zeros of output type. Others are finitely generated, but free; and these have an analogous classification as zeros of input type. It is possible to extend the idea of zeros in such a way as to embrace both the zeros of pole-type, as in the previous section, and zeros of output or input type. The development of these notions, as well as their interrelationships, is the goal of this section.

3.1. The Γ-zero module. We begin our discussion of extended zeros in such a way as to address the zeros of output type. Consider the $k[z]$-factor module

$$Z_\Gamma(M(z)) = M^{-1}U[z]/\{R[z] \cap M^{-1}U[z]\}. \tag{3.1}$$

This module is torsion. Indeed, observe that $M^{-1}U[z]$ is a submodule of $R(z)$, from which we have for every representative $r(z)$ of a class in (3.1) a polynomial $p(z)$ in $k[z]$ such that $p(z)r(z)$ is in $R[z]$ and is therefore equivalent to zero. If $\ker M(z)$ is not equal to zero, however, (3.1) is not finitely generated. We shall call (3.1) the Γ-zero module of $M(z)$.

When $\ker M(z)$ is zero, $Z_\Gamma(M(z))$ is identical to the form (2.8) and so is isomorphic, as a $k[z]$-module, to the zero module $Z(M(z))$ of § 2. Note that this statement holds whether or not $M(z)$ has a nontrivial cokernel. When $\ker M(z)$ is not zero, the nature of the Γ-zero module can be studied by means of the commutative diagram in Fig. 3.1, where rows and columns are natural inclusions. By the action

$$r(z) \mapsto r(z) \bmod (\ker M(z) + R[z]) \cap M^{-1}U[z], \tag{3.2}$$

this diagram induces an epic, $k[z]$-linear map $\alpha(z)$ from $Z_\Gamma(M(z))$ onto the isomorphic copy (2.8) of $Z(M(z))$. The kernel of $\alpha(z)$ is isomorphic to

$$\Gamma(M(z)) = \ker M(z)/\{R[z] \cap \ker M(z)\}. \tag{3.3}$$

Thus we have the short exact sequence

$$0 \to \Gamma(M(z)) \to Z_\Gamma(M(z)) \xrightarrow{\alpha(z)} Z(M(z)) \to 0, \tag{3.4}$$

if we identify (2.8) with $Z(M(z))$.

The module (3.3) is both torsion and divisible. To see the latter, observe that scalar multiplication $p(z): \ker M(z) \to \ker M(z)$ is an epimorphism of $k[z]$-modules whenever $p(z)$ is nonzero in $k[z]$. Then $\Gamma(M(z))$ inherits this same property. As a divisible module over $k[z]$, $\Gamma(M(z))$ is injective; and so its image under injection into

$$\begin{tikzcd}
R[z] \cap M^{-1}U[z] \arrow[r] \arrow[d] & M^{-1}U[z] \arrow[d] \\
(\ker M(z) + R[z]) \cap M^{-1}U[z] \arrow[r] & M^{-1}U[z]
\end{tikzcd}$$

FIG. 3.1

$Z_\Gamma(M(z))$ is a direct summand. It follows that

$$Z_\Gamma(M(z)) \approx \Gamma(M(z)) \oplus Z(M(z)). \tag{3.5}$$

We will refer to $\Gamma(M(z))$ as the divisible zero module of $M(z)$, and to $Z_\Gamma(M(z))$, the Γ-zero module of $M(z)$, as an extended zero module. In view of (3.5), we see that the notion of Γ-extended zeros includes both $\Gamma(M(z))$ and $Z(M(z))$ as special cases.

Realization theory also makes use of a torsion, divisible module. Indeed, let

$$M^{\#}(z) = pM(z)i \tag{3.6}$$

be the $k[z]$-linear composition of the natural inclusion $i: R[z] \to R(z)$, $M(z)$, and the natural projection $p: U(z) \to U(z)/U[z]$. $M^{\#}(z)$ is the restricted input/output map corresponding to $M(z)$. The output module of $M^{\#}(z)$ is torsion divisible, and is often denoted by ΓU. From the commutative diagram of Fig. 3.2, of natural inclusions, we can induce a monic, $k[z]$-linear map on $Z_\Gamma(M(z))$ into ΓR. Then we can think of $Z_\Gamma(M(z))$ as a submodule of ΓR. Accordingly, we employ the subscript Γ and regard Γ-zeros as being of output type.

3.2. The Ω-zero module. Next we examine zeros of input type, by forming the $k[z]$-factor module

$$Z_\Omega(M(z)) = U[z]/\{U[z] \cap MR[z]\}. \tag{3.7}$$

Module (3.7) is finitely generated, because $U[z]$ is finitely generated. Not every equivalence class in (3.7) is torsion, however, because not every $u(z)$ in $U[z]$ can be brought into $MR[z]$ by means of scalar multiplication. If the image, im $M(z)$, of $M(z)$ is equal to $U(z)$ on the $k(z)$-level, then the isomorphic form (2.9) of $Z(M(z))$ is equal to (3.7); and so we see that $Z_\Omega(M(z))$ differs from $Z(M(z))$ when coker $M(z)$ is nonzero. We call (3.7) the Ω-zero module of $M(z)$.

$$\begin{tikzcd}
R[z] \cap M^{-1}U[z] \arrow[r] \arrow[d] & M^{-1}U[z] \arrow[d] \\
R[z] \arrow[r] & R(z)
\end{tikzcd}$$

FIG. 3.2

When coker $M(z)$ is nonzero, we observe that form (2.9) is a submodule of (3.7). Identifying (2.9) with $Z(M(z))$, we have a natural inclusion of $Z(M(z))$ into $Z_\Omega(M(z))$, with cokernel isomorphic as a $k[z]$-module to

$$\Omega(M(z)) = U[z]/\{U[z] \cap MR(z)\}. \tag{3.8}$$

Thus there is a short exact sequence

$$0 \to Z(M(z)) \to Z_\Omega(M(z)) \to \Omega(M(z)) \to 0 \tag{3.9}$$

of $k[z]$-modules and $k[z]$-linear maps. The module (3.8) is torsion-free. To determine this, suppose the contrary. Then there is a nonzero $u(z)$ in $U[z]$ and a nonzero $p(z)$ in $k[z]$ such that

$$p(z)u(z) = M(z)r(z), \qquad r(z) \in R(z). \tag{3.10}$$

But then it follows that $u(z)$ is in $M(z)R(z)$ and must be equivalent to zero in (3.8). As a finitely generated, torsion-free module over the principal ideal domain $k[z]$,

$\Omega(M(z))$ is a free module. From the exactness of (3.9) at $\Omega(M(z))$, it then follows that

$$Z_\Omega(M(z)) \approx \Omega(M(z)) \oplus Z(M(z)). \tag{3.11}$$

We will call $\Omega(M(z))$ the free zero module of $M(z)$. Like $Z_\Gamma(M(z))$, then, $Z_\Omega(M(z))$ is an extended zero module, this time of Ω-type. Note from (3.11) that extended Ω-zeros include $\Omega(M(z))$ and $Z(M(z))$ as particular cases.

From the viewpoint of realization theory, the inputs to $M^{\#}(z)$ lie in $R[z]$, which is often denoted ΩR. As a factor module of ΩU, $Z_\Omega(M(z))$ may be thought of in terms of zeros of input type; and this motivates the subscript Ω.

3.3. Relation between $Z_\Gamma(M(z))$ and $Z_\Omega(M(z))$. The presence of $Z(M(z))$ in both the Γ-zero module, as a factor module, and the Ω-zero module, as a submodule, suggests that there may exist a relationship between Γ-zeros and Ω-zeros. Indeed, this is the case; and a convenient way in which to state this relationship is provided in the following theorem.

THEOREM 1. *Let $Z_\Gamma(M(z))$ be the Γ-zero module of $M(z)$, and let $Z_\Omega(M(z))$ be its Ω-zero module. Then there exists an exact sequence*

$$0 \to \Gamma(M(z)) \to Z_\Gamma(M(z)) \to Z_\Omega(M(z)) \to \Omega(M(z)) \to 0 \tag{3.12}$$

of $k[z]$-modules and $k[z]$-linear maps.

Proof. We establish a $k[z]$-linear map from $Z_\Gamma(M(z))$ into $Z_\Omega(M(z))$ by inducing from the restriction of $M(z)$ to $M^{-1}U[z]$. The existence of the desired map is a consequence of the commutative diagram of Fig. 3.3, where rows are natural inclusions and columns are restrictions of $M(z)$. The kernel of the induced map is given by

$$M^{-1}(MR[z] \cap U[z])/\{R[z] \cap M^{-1}U[z]\}, \tag{3.13}$$

which is isomorphic as a $k[z]$-module to $\Gamma(M(z))$. The image of the induced map is form (2.9); and so the cokernel is $\Omega(M(z))$, as asserted.

$$\begin{tikzcd}
R[z] \cap M^{-1}U[z] \arrow[r] \arrow[d, "M(z)"'] & M^{-1}U[z] \arrow[d, "M(z)"] \\
MR[z] \cap U[z] \arrow[r] & U[z]
\end{tikzcd}$$

FIG. 3.3

Remark. It is interesting to consider what happens in Theorem 1, with respect to its exact sequence (3.12), when the $k(z)$-linear map $M(z)$ has special features. For example, when $\ker M(z)$ is zero, so that its matrix representation would have full column rank over $k(z)$, then (3.3) indicates that $\Gamma(M(z))$ is zero and (3.5) shows that $Z_\Gamma(M(z))$ is isomorphic as a $k[z]$-module to $Z(M(z))$. Thus this special case of (3.12) becomes identical to the corresponding special case of (3.9). In like fashion, if the matrix of $M(z)$ has full row rank over $k(z)$, then $\Omega(M(z))$ becomes zero; and the corresponding specialization of (3.12) is identical to the associated specialization of (3.4).

Remark. It should be noted that $Z_\Gamma(M(z))$ and $Z_\Omega(M(z))$ will not be isomorphic unless the matrix of $M(z)$ is both square and invertible. For instance, it is quite possible for $Z_\Omega(M(z))$ to have finite dimension as a vector space over k while $Z_\Gamma(M(z))$ is infinite-dimensional. This occurs when the matrix of $M(z)$ is right invertible over $k(z)$, but not invertible.

In this section, we have discussed extensions of the classical idea of zero to include divisible or free zeros. These extensions were expressed, respectively, in terms of the

Γ-zero module $Z_\Gamma(M(z))$ and the Ω-zero module $Z_\Omega(M(z))$ of a $k(z)$-linear map $M(z): R(z) \to U(z)$. The Γ-zero module has received prior study in [14], where it has been shown to be closely related to the concept of zero signal generation. The Ω-zero module does not seem to have received prior attention. Both Γ-zeros and Ω-zeros play a central role in studying fixed zeros in model matching problems. We call such fixed modules "matching zeros," and discuss them in the following section.

4. Matching zeros. Suppose that the dimensions of R, U, and Y are each equal to unity, and consider the scalar equation

$$t(z) = p(z)m(z), \tag{4.1}$$

with all entries in $k(z)$. If we express these transfer functions in relatively prime form

$$t(z) = (n_t(z), d_t(z)), p(z) = (n_p(z), d_p(z)), m(z) = (n_m(z), d_m(z)), \tag{4.2}$$

then we obtain

$$p(z) = (n_t(z)d_m(z), d_t(z)n_m(z)), \tag{4.3}$$

which is useful when $m(z)$ is given, and

$$m(z) = (n_t(z)d_p(z), d_t(z)n_p(z)), \tag{4.4}$$

which is useful when $p(z)$ is given. From (4.3), we see that the zeros of $p(z)$ consist of those zeros of $t(z)$ which are not zeros of $m(z)$, together with those poles of $m(z)$ which are not poles of $t(z)$. From (4.4), we have that the zeros of $m(z)$ are zeros of $t(z)$ which are not zeros of $p(z)$, together with poles of $p(z)$ which are not poles of $t(z)$. This idea for determining fixed zeros in solutions to (4.1) is direct and intuitive; and we may well inquire if there exists a suitable extension to the multivariable case. Unfortunately, as we will later demonstrate in detail, such a generalization is not really available when we restrict our attention to $k[z]$-zeros which are finitely generated and torsion. But, if we permit the use of extended zeros, as explained in § 3, then we can obtain statements which are quite close in spirit to those above.

In order to capture in a more precise way the intuitive ideas of the preceding paragraph, we introduce the notion of a matching zero module. The matching Γ-zero module is denoted by Z_Γ and defined to be the $k[z]$-factor module

$$Z_\Gamma = T^{-1}Y[z]/\{T^{-1}Y[z] \cap M^{-1}U[z]\}, \tag{4.5}$$

where $T(z): R(z) \to Y(z)$ is a $k(z)$-linear map. The matching Ω-zero module is denoted by Z_Ω and defined to be the $k[z]$-factor module

$$Z_\Omega = \{PU[z] + TR[z]\}/TR[z], \tag{4.6}$$

for $P(z): U(z) \to Y(z)$ a $k(z)$-linear map. The algebraic character of matching zero modules of Γ-type and of Ω-type is settled by the theorem which follows.

THEOREM 2. *Let Z_Γ and Z_Ω be the matching zero modules of Γ-type and Ω-type, respectively, for $k(z)$-linear maps $M(z): R(z) \to U(z)$, $P(z): U(z) \to Y(z)$, and $T(z): R(z) \to Y(z)$. Define $k[z]$-factor modules*

$$Z'_\Gamma = T^{-1}Y[z]/\{T^{-1}Y[z] \cap (\ker T(z) + M^{-1}U[z])\}, \tag{4.7}$$

$$\Gamma = \ker T(z)/\{\ker T(z) \cap M^{-1}U[z]\}, \tag{4.8}$$

$$Z'_\Omega = \{PU[z] \cap TR(z)\}/\{PU[z] \cap TR[z]\}, \tag{4.9}$$

$$\Omega = PU[z]/\{PU[z] \cap TR(z)\}. \tag{4.10}$$

Then

$$Z_\Gamma \approx \Gamma \oplus Z'_\Gamma, \tag{4.11}$$

where Γ *is torsion divisible and* Z'_Γ *is finitely generated and torsion*; *and*

$$Z_\Omega \approx \Omega \oplus Z'_\Omega, \tag{4.12}$$

where Ω *is finitely generated and free while* Z'_Ω *is finitely generated and torsion.*

Proof. We begin by discussing the matching Ω-zero module. Because $R[z]$ and $U[z]$ are finitely generated, we have also from (4.6) that Z_Ω bears the same property. However, not every element in Z_Ω is a torsion element. Consider the torsion submodule of Z_Ω. If $y(z)$ represents a torsion element, then there exists a nonzero polynomial $p(z)$ such that

$$p(z)y(z) = T(z)r(z), \qquad r(z) \in R[z]. \tag{4.13}$$

But (4.13) implies that $y(z) \in T(z)R(z)$. Conversely, if $y(z) \in T(z)R(z)$, then (4.13) can be satisfied for appropriate choices of $p(z)$ and $r(z)$. Thus the torsion submodule of Z_Ω is given by

$$[\{PU[z]+TR[z]\} \cap TR(z)]/TR[z] = \{PU[z] \cap TR(z) + TR[z]\}/TR[z] \tag{4.14a}$$

$$\approx \{PU[z] \cap TR(z)\}/\{PU[z] \cap TR[z]\} \tag{4.14b}$$

$$= Z'_\Omega. \tag{4.14c}$$

Using the isomorphic form (4.14a), we can insert Z'_Ω into Z_Ω and produce a cokernel which is isomorphic to

$$\begin{aligned}&\{PU[z]+TR[z]\}/\{PU[z] \cap TR(z) + TR[z]\}\\ &\quad\approx PU[z]/\{PU[z] \cap TR(z)\} = \Omega.\end{aligned} \tag{4.15}$$

We claim that Ω is a torsion-free module. To see this, suppose that $y(z)$ represents a torsion element; then

$$p(z)y(z) = T(z)r(z), \qquad r(z) \in R(z), \tag{4.16a}$$

for some nonzero polynomial $p(z)$. Consequently, there is a polynomial $q(z)$ such that

$$(q(z)p(z))y(z) = T(z)(q(z)r(z)), \qquad q(z)r(z) \in R[z]. \tag{4.16b}$$

But this means that $y(z)$ is equivalent to zero in Z_Ω, as desired. As a finitely generated module over a principal ideal domain, Ω is also free. We thus have a short exact sequence

$$0 \to Z'_\Omega \to Z_\Omega \to \Omega \to 0; \tag{4.17}$$

and, because Ω is free, and a factor module of Z_Ω, we obtain (4.12). Next we turn to the matching Γ-zero module. If $r(z)$ is an element of $T^{-1}Y[z]$, and if the equation $p(z)x(z) = r(z)$ has a solution $x(z)$ in $T^{-1}Y[z]$ for each $p(z)$ in $k[z]$, then

$$T(z)(r(z)/p(z)) \in Y[z] \tag{4.18}$$

for all nonzero $p(z)$. But this cannot be true unless $r(z) \in \ker T(z)$. It follows that $\ker T(z)$ is the divisible submodule of $T^{-1}Y[z]$. Consequently, Γ is divisible and is isomorphic to the divisible submodule of Z_Γ. There exists an epic, $k[z]$-linear map from Z_Γ onto Z'_Γ, with kernel isomorphic to Γ. Thus we can write a short exact sequence

$$0 \to \Gamma \to Z_\Gamma \to Z'_\Gamma \to 0, \tag{4.19}$$

and infer (4.11) from the exactness of (4.19) at Γ, together with the fact that Γ is divisible. The torsion character of Γ and Z'_Γ follows, of course, from that of Z_Γ; and the torsion character of Z_Γ is a result of the fact that

$$M(T^{-1}Y[z]) \subset U(z). \tag{4.20}$$

Remark. It is evident from Theorem 2 that Z_Γ is a module having the same character as a module of extended Γ-zeros. In the same way, Z_Ω has a structure of the same type as a module of extended Ω-zeros.

Remark. Observe that Γ in (4.8) vanishes when the matrix of $T(z)$ has full column rank over $k(z)$. In this case the matching zero module of Γ-type is finitely generated and torsion. In similar fashion, we can say that Ω in (4.10) becomes zero when the matrix of $T(z)$ has full row rank over $k(z)$. Then the matching zero module of Ω-type is finitely generated and torsion.

Remark. It is worth pointing out that a $k(z)$-linear map, such as $T(z)$, may be onto while its restriction to polynomial vectors may not. It is precisely this fact which makes the study of zero properties possible for matrices. Thus, when we speak of the row rank or the column rank of a matrix, it is important to make clear whether we are regarding it as a linear map over $k[z]$ or $k(z)$.

So as to justify the use of the term "matching zeros" in connection with the modules Z_Γ and Z_Ω, we will show that they appear naturally as submodules and factor modules of the appropriate extended zero modules associated with solutions $M(z)$ or $P(z)$ to the model matching equation of § 1.

THEOREM 3. *Suppose that Z_Γ and Z_Ω are the matching zero modules of Γ-type and Ω-type, respectively, and let $T(z): R(z) \to Y(z)$ be a $k(z)$-linear map.*

(1) *If $P(z): U(z) \to Y(z)$ is a $k(z)$-linear map whose image contains that of $T(z)$, and if $M(z): R(z) \to U(z)$ is a $k(z)$-linear map which satisfies the equation $T(z) = P(z)M(z)$, then there exists an epic, $k[z]$-linear map*

$$\beta_\Omega(z): Z_\Omega(M(z)) \to Z_\Omega, \tag{4.21a}$$

so that Z_Ω is isomorphic to a factor module of the Ω-zero module of $M(z)$.

(2) *If $M(z): R(z) \to U(z)$ is a $k(z)$-linear map whose kernel is contained in that of $T(z)$, and if $P(z): U(z) \to Y(z)$ is a $k(z)$-linear map which satisfies the equation $T(z) = P(z)M(z)$, then there exists a monic, $k[z]$-linear map*

$$\beta_\Gamma(z): Z_\Gamma \to Z_\Gamma(P(z)), \tag{4.21b}$$

so that Z_Γ is isomorphic to a submodule of the Γ-zero module of $P(z)$.

Discussion of Theorem 3. From (4.5), Z_Γ depends upon $T(z)$ and $M(z)$, but not upon $P(z)$; accordingly, the constraint indicated by (4.21b) holds for *all* solutions $P(z)$ to the model matching equation. Likewise, (4.6) shows that Z_Ω depends only upon $P(z)$ and $T(z)$; and so (4.21a) is a feature of *all* solutions $M(z)$.

Proof of Theorem 3. Recall from § 3.2 the definition (3.7) for the Ω-zero module of $M(z)$. From the commutative diagram of Fig. 4.1, which is made up of natural inclusions and restrictions of $P(z)$, we can induce the map $\beta_\Omega(z)$, which is epic because the restriction of column two is epic. We remark that row two of Fig. 4.1 induces an isomorphic copy of Z_Ω, which is employed for convenience. Next, consider the

$$\begin{tikzcd}
U[z] \cap MR[z] \arrow[r] \arrow[d, "P(z)"'] & U[z] \arrow[d, "P(z)"] \\
PU[z] \cap TR[z] \arrow[r] & PU[z]
\end{tikzcd}$$

FIG. 4.1

commutative diagram of Fig. 4.2. The second row of the diagram induces $Z_\Gamma(P(z))$, in accord with definition (3.1), § 3.1, while the first row induces Z_Γ, as given in (4.5). Columns one and two are restrictions of $M(z)$. Map $\beta_\Gamma(z)$ of the theorem can be induced from this diagram. The calculation of its kernel gives

$$\{M^{-1}(P^{-1}Y[z]\cap U[z])\cap T^{-1}Y[z]\}/\{T^{-1}Y[z]\cap M^{-1}U[z]\}, \tag{4.22}$$

which vanishes as desired.

Remark. Theorem 3 has a number of features which occur in special cases. From Theorem 2, for example, we know that Z_Ω becomes a finitely generated, torsion module when $T(z)$ is an epic $k(z)$-linear map. The question arises about whether the solution $M(z)$ of (4.21a) will have extended zeros of Ω-type. It is easy to see that any nonzero element of $Z_\Omega(M(z))$ must either be a torsion element or lie in the kernel of $\beta_\Omega(z)$. But we know that the module $\Omega(M(z))$ contains no nonzero torsion elements. Thus, the question of whether a solution $M(z)$ has zeros of Ω-type can be related to whether $\beta_\Omega(z)$ has a kernel. We will see in § 7 that this kernel can be made to vanish by requiring that the kernel of $P(z)$ be zero. Accordingly, if the matrices of $T(z)$ and $P(z)$, considered as $k(z)$-linear maps, are of full row rank and column rank, respectively, then the unique solution $M(z)$ has no extended zeros of Ω-type. Moreover, the finitely generated, torsion zeros of $M(z)$ arise from the resulting, finite-dimensional matching zero module Z_Ω. Related remarks can be made if $T(z)$ is a monic $k(z)$-linear map. In this case Z_Γ is finitely generated and torsion; and the center of attention is the cokernel of $\beta_\Gamma(z)$. The extended Γ-zero character of solutions $P(z)$ can be eliminated by assuming that $M(z)$ has zero cokernel. Thus, if the matrices of $T(z)$ and $M(z)$, considered as $k(z)$-linear maps, are of full column rank and row rank, respectively, then the unique solution $P(z)$ has no extended zeros of Γ-type. For brevity, we leave the details of this second argument to the reader.

The constructions of (4.3) and (4.4) constitute a direct and intuitive way to characterize the zeros of solutions $p(z)$ and $m(z)$ to the model matching equation (4.1). In a generalization of those ideas to cases in which the dimensions of R, U, and Y are not equal to unity, we have replaced the elementary notion of polynomial with that of the more powerful and versatile module. One basic reason why we advance this point of view has to do with the fact that solutions $P(z)$ or $M(z)$ need not be unique, as they are in (4.1). We have to deal, then, with the module of zeros which is somehow a part of the zero module of any possible solution. The matching zero modules Z_Γ and Z_Ω play this role, for the two possible cases; and containment has to be understood either in terms of a submodule or in terms of a factor module.

However, an even more fundamental reason for introducing the matching zero modules has to do with generalizing the interpretations of (4.3) and (4.4), which follow those equations. It turns out that the analogous statements for higher-dimensional R, U, and Y must be made in terms of extended zeros. Because extended zeros are transparent in the familiar polynomial-matrix theories, it becomes both necessary and

$$\begin{tikzcd}
T^{-1}Y[z]\cap M^{-1}U[z] \arrow[r] \arrow[d, "M(z)"'] & T^{-1}Y[z] \arrow[d, "M(z)"] \\
P^{-1}Y[z]\cap U[z] \arrow[r] & P^{-1}Y[z]
\end{tikzcd}$$

FIG. 4.2

prudent to make use of the module-theoretic method in order to give them a careful accounting. The section following provides further detail.

5. Significance of Z_Γ and Z_Ω. We have seen in § 4 that the matching zero module Z_Γ is isomorphic to a direct sum of divisible zeros with zeros of finitely generated, torsion type. From § 2, we see that the latter type of module appears both in representations of poles and in representation of zeros. The former type of module, however, has appeared in § 3, which deals with extended zeros. In an analogous way, the matching zero module Z_Ω is isomorphic to a direct sum of free zeros with zeros of finitely generated, torsion type. Once again, the second class of module appears also in descriptions of poles, while the first type has been employed to discuss extended zeros. If, therefore, we are to generalize the conclusions which follow (4.3) and (4.4), we may anticipate that the term "zeros" will be replaced by the term "extended zeros." We might also hope to represent Z_Γ in terms of Γ-zeros of $T(z)$ and of $M(z)$, together with poles of $T(z)$ and of $M(z)$, and to represent Z_Ω with Ω-zeros of $T(z)$ and of $P(z)$, as well as with poles of $T(z)$ and of $P(z)$. These intuitions turn out to be true in great part.

The difference which arises has to do with the treatment of $M(z)$, in a discussion of Z_Γ, and with $P(z)$, in a treatment of Z_Ω. Basically, we must pair these maps with $T(z)$, in place of letting them stand alone. Thus, for Z_Γ, we study $T(z)$ and the pair $(T(z), M(z))$; for Z_Ω, we take $T(z)$ and the pair $(T(z), P(z))$. This section shows just what conceptual modifications have to be made. The main result is Theorem 4, following.

THEOREM 4. *Let Z_Γ and Z_Ω be the matching zero modules of Γ-type and Ω-type, respectively, for $k(z)$-linear maps $M(z): R(z) \to U(z)$, $P(z): U(z) \to Y(z)$, and $T(z): R(z) \to Y(z)$.*

(1) *If $[T(z)\ P(z)]: R(z) \oplus U(z) \to Y(z)$ is the $k(z)$-linear map with action*

$$[T(z)\ P(z)](r(z), u(z)) = T(z)r(z) + P(z)u(z), \tag{5.1}$$

then there exist $k[z]$-modules Z_1 and P_1, together with appropriate $k[z]$-linear maps, such that the following three short sequences are exact:

$$0 \to P(T(z)) \to P([T(z)\ P(z)]) \to P_1 \to 0, \tag{5.2a}$$

$$0 \to Z_1 \to Z_\Omega(T(z)) \to Z_\Omega([T(z)\ P(z)]) \to 0, \tag{5.2b}$$

$$0 \to Z_1 \to Z_\Omega \to P_1 \to 0. \tag{5.2c}$$

(2) *If*

$$\begin{bmatrix} T(z) \\ M(z) \end{bmatrix} : R(z) \to Y(z) \oplus U(z) \tag{5.3a}$$

is the $k(z)$-linear map having action

$$\begin{bmatrix} T(z) \\ M(z) \end{bmatrix} (r(z)) = (T(z)r(z), M(z)r(z)), \tag{5.3b}$$

then there exist $k[z]$-modules Z_2 and P_2, together with appropriate $k[z]$-linear maps, such that the following three short sequences are exact:

$$0 \to P_2 \to P\left(\begin{bmatrix} T(z) \\ M(z) \end{bmatrix}\right) \to P(T(z)) \to 0, \tag{5.4a}$$

$$0 \to Z_\Gamma\left(\begin{bmatrix} T(z) \\ M(z) \end{bmatrix}\right) \to Z_\Gamma(T(z)) \to Z_2 \to 0, \tag{5.4b}$$

$$0 \to P_2 \to Z_\Gamma \to Z_2 \to 0. \tag{5.4c}$$

Discussion of Theorem 4. The module P_1 is described in (5.2a), on a module-theoretic level, as "the poles of $[T(z)\ P(z)]$ which are not poles of $T(z)$." Note that, in the case (4.4), this reduces to the poles of $P(z)$ which are not poles of $T(z)$, as desired. The pole module of $[T(z)\ P(z)]$ has generalized "the poles of $P(z)$" for this situation. In the same way, from (5.2b), we see that the Ω-zero module of $T(z)$ has generalized "the zeros of $T(z)$," while the Ω-zero module of $[T(z)\ P(z)]$ has generalized "the zeros of $P(z)$." Note that, in the situation (4.4), $Z_\Omega(T(z)) = Z(T(z))$, and $Z_\Omega([T(z)\ P(z)]) = Z([T(z)\ P(z)])$. With the nature of Z_1 and P_1 established, (5.2c) shows that Z_Ω has a submodule Z_1, and a factor module P_1 containing those zeros of Z_Ω which are not zeros of Z_1. The explanation of (5.4) proceeds in a corresponding manner. Relative to (4.3), the Γ-zero module of $T(z)$ replaces "the zeros of $T(z)$," while the Γ-zero module of (5.3) replaces "the zeros of $M(z)$." The pole modules of $T(z)$ and of (5.3) replace "the poles of $T(z)$" and "the poles of $M(z)$," respectively. Once again, in the case (4.3), $Z_\Gamma(T(z)) = Z(T(z))$, and so forth.

Remark. In (5.2b) and (5.2c), it is apparent that none of the modules in the sequence will be extended of Ω-type if we assume that $T(z)$ is epic as a $k(z)$-linear map. This has already been discussed for $Z_\Omega(T(z))$ and Z_Ω. It follows for the right member in (5.2b) because $[T(z)\ P(z)]$ is onto when $T(z)$ is onto, and for the left member in (5.2b) and (5.2c) because Z_1 is a submodule in each case. Similar comments follow in (5.4b) and (5.4c) when $T(z)$ is monic as a $k(z)$-linear map. For these cases, Theorem 4 is for zeros a result corresponding to that obtained for fixed poles by Conte, Perdon, and Wyman [7]. The work in [7], however, required only the concepts of finitely generated, torsion poles and zeros. As we will see in §§ 7 and 8, the intuitively pleasing character of Theorem 4 is not carried over to the features of finite-dimensional zeros.

Proof of Theorem 4. We begin with the first part of the theorem and form a commutative diagram of natural inclusions, as in Fig. 5.1. Observe that column two has a cokernel which is Z_Ω, as in (4.6). Next consider the $k(z)$-linear map

$$[T(z)\ P(z)]: R(z) \oplus U(z) \to Y(z), \tag{5.5}$$

with action (5.1). The pole module of the map (5.5) is given by the isomorphic form (2.5) to be

$$P([T(z)\ P(z)]) = \frac{[T\,P](R[z] \oplus U[z])}{Y[z] \cap [T\,P](R[z] \oplus U[z])}, \tag{5.6}$$

which we recognize as the quotient module induced from row two of Fig. 5.1. Row one, of the same diagram, on the other hand, induces a quotient module

$$\frac{TR[z]}{Y[z] \cap TR[z]} = P(T(z)), \tag{5.7}$$

the pole module of $T(z)$. The diagram, moreover, induces a monic $k[z]$-linear map on $P(T(z))$ into $P([T(z)\ P(z)])$, with a cokernel that we have denoted by P_1. Return

$$\begin{tikzcd}
Y[z] \cap TR[z] \arrow[r] \arrow[d] & TR[z] \arrow[d] \\
Y[z] \cap \{PU[z] + TR[z]\} \arrow[r] & PU[z] + TR[z]
\end{tikzcd}$$

FIG. 5.1

now to the composite map (5.5), and form the Ω-zero module

$$Z_\Omega([T(z)\ \ P(z)]) = \frac{Y[z]}{Y[z] \cap [T\ P](R[z] \oplus U[z])}. \tag{5.8}$$

Next calculate the Ω-zero module of $T(z)$ in the manner

$$Z_\Omega(T(z)) = \frac{Y[z]}{Y[z] \cap TR[z]}. \tag{5.9}$$

By the inclusion $TR[z] \subset TR[z] + PU[z]$, there is an epic, $k[z]$-linear map on $Z_\Omega(T(z))$ onto $Z_\Omega([T(z)\ \ P(z)])$, with kernel isomorphic to

$$\frac{Y[z] \cap (TR[z] + PU[z])}{Y[z] \cap TR[z]}. \tag{5.10}$$

In Fig. 5.1, however, see that (5.10) is just the cokernel of column one, which we have denoted by Z_1. Equation (5.2c) then follows. Next we consider the second part of the theorem. A useful beginning point is the commutative diagram of Fig. 5.2, in which once again both the rows and the columns are natural inclusions. From the second row, we can induce $Z_\Gamma(T(z))$, while from the first row we obtain

$$Z_\Gamma\left(\begin{bmatrix} T(z) \\ M(z) \end{bmatrix}\right) \tag{5.11}$$

by the analogous process. The diagram of Fig. 5.2 then induces a monic, $k[z]$-linear map on the module (5.11) into $Z_\Gamma(T(z))$, with cokernel Z_2 as in (5.4b). The cokernel of column two is Z_Γ, by the definition (4.5). Because the cokernels of columns one and two, together with that of the induced monomorphism, fit into a short exact sequence, we have (5.4c) if P_2 is defined to be the cokernel of column one. This last cokernel can be related to certain pole modules. Recall definition (2.3), and form the pair of pole modules

$$P\left(\begin{bmatrix} T(z) \\ M(z) \end{bmatrix}\right) = \frac{R[z]}{R[z] \cap T^{-1}Y[z] \cap M^{-1}U[z]}, \tag{5.12}$$

$$P(T(z)) = \frac{R[z]}{R[z] \cap T^{-1}Y[z]}. \tag{5.13}$$

Clearly, $R[z] \cap T^{-1}Y[z] \cap M^{-1}U[z]$ is a submodule of $R[z] \cap T^{-1}Y[z]$, and so there exists an epic, $k[z]$-linear map from the module of (5.12) onto the module of (5.13). The kernel of this epimorphism is isomorphic as a $k[z]$-module to

$$\frac{R[z] \cap T^{-1}Y[z]}{R[z] \cap T^{-1}Y[z] \cap M^{-1}U[z]}, \tag{5.14}$$

which is P_2 in (5.4a).

$$\begin{tikzcd}
T^{-1}Y[z] \cap M^{-1}U[z] \cap R[z] \arrow[r] \arrow[d] & T^{-1}Y[z] \cap M^{-1}U[z] \arrow[d] \\
T^{-1}Y[z] \cap R[z] \arrow[r] & T^{-1}Y[z]
\end{tikzcd}$$

FIG. 5.2

Because the pole modules of (5.3) and of (5.5) are finitely generated and torsion, we have from the exactness of (5.2a) at P_1 and from the exactness of (5.4a) at P_2 that these modules have the same character. On the other hand, Z_1 and Z_2 may display behavior which is in part free or divisible, respectively. We conclude this section by giving an explicit description of these features.

Consider the module Z_1, as given in (5.10). Finitely generated, Z_1 has a torsion submodule. Suppose that $y(z)$ in $Y[z]$ represents a torsion element. Then there is a nonzero polynomial $p(z)$ in $k[z]$ such that $p(z)y(z)$ is in $TR[z]$, which means that $y(z)$ lies in $TR(z)$. The converse is also true, and so

$$Y[z]\cap(TR[z]+PU[z])\cap TR(z)=Y[z]\cap\{TR[z]+TR(z)\cap PU[z]\} \tag{5.15}$$

represents the torsion submodule Z_1' of Z_1. We can then set up a short exact sequence

$$0\to Z_1'\to Z_1\to\Omega_1\to 0, \tag{5.16}$$

in which Ω_1 may be taken as

$$\frac{Y[z]\cap(TR[z]+PU[z])}{Y[z]\cap(TR[z]+PU[z])\cap TR(z)}. \tag{5.17}$$

Now Ω_1 is torsion-free, because any representative which can be annihilated by a polynomial must already stand for the zero element. It then follows that Z_1 is isomorphic to a direct sum of Ω_1 and Z_1'. Next, we refer back to Fig. 5.2, where the cokernel of the induced monomorphism has been defined to be Z_2. Note that the image of the monomorphism is

$$\frac{T^{-1}Y[z]\cap M^{-1}U[z]+T^{-1}Y[z]\cap R[z]}{T^{-1}Y[z]\cap R[z]}, \tag{5.18}$$

so that we have the form

$$T^{-1}Y[z]/\{T^{-1}Y[z]\cap M^{-1}U[z]+T^{-1}Y[z]\cap R[z]\} \tag{5.19}$$

for Z_2. A torsion module, (5.19) has a divisible submodule given by

$$\Gamma_2=\ker T(z)/\{\ker T(z)\cap\{T^{-1}Y[z]\cap M^{-1}U[z]+T^{-1}Y[z]\cap R[z]\}\}. \tag{5.20}$$

If we establish a short exact sequence

$$0\to\Gamma_2\to Z_2\to Z_2'\to 0, \tag{5.21}$$

then Z_2' can be taken to be

$$T^{-1}Y[z]/\{\ker T(z)+T^{-1}Y[z]\cap M^{-1}U[z]+T^{-1}Y[z]\cap R[z]\}, \tag{5.22}$$

which is finitely generated. We summarize the discussion in the next theorem.

THEOREM 5. *Let Z_1 be the extended zero module of Ω-zeros of $T(z)$ which are not Ω-zeros of* (5.5), *and let Z_2 be the extended zero module of Γ-zeros of $T(z)$ which are not Γ-zeros of* (5.3). *Then there exist $k[z]$-modules Ω_1 and Z_1', Γ_2, and Z_2', such that*

$$Z_1\approx\Omega_1\oplus Z_1', \tag{5.23a}$$

$$Z_2\approx\Gamma_2\oplus Z_2', \tag{5.23b}$$

where Z_i', $i=1,2$, are finitely generated, torsion modules, Ω_1 is finitely generated and free, and Γ_2 is torsion divisible.

Proof. The details have been provided in the prelude to the theorem. Z_1' and Z_2' were discussed in (5.15) and (5.22); Ω_1 and Γ_2 are given by (5.17) and (5.20).

Remark. As remarked after the preceding theorem, Ω_1 or Γ_2 vanish if $T(z)$, as a $k(z)$-linear map, is epic or monic, respectively.

6. Essential solutions. We have seen in Theorem 3 that the modules Z_Γ and Z_Ω of matching zeros represent zero constraints which must hold for all solutions to the model matching equations. In particular, suppose that $T(z): R(z) \rightarrow Y(z)$ is a $k(z)$-linear map. When $P(z): U(z) \rightarrow Y(z)$ is a $k(z)$-linear map whose image contains that of $T(z)$, then Z_Ω is a factor module of the Ω-zero module of every $k(z)$-linear solution $M(z): R(z) \rightarrow U(z)$ to the equation $T(z) = P(z)M(z)$. Or, if we regard $M(z)$ to be given, with kernel included in that of $T(z)$, then Z_Γ must be a submodule of the Γ-zero module of every solution $P(z)$. The mechanism for the constraint is embodied in the $k[z]$-linear maps $\beta_\Omega(z)$ and $\beta_\Gamma(z)$ of (4.21). In this section, we examine initially the conditions under which the solution may have an extended zero module isomorphic to a matching zero module. This question is, of course, fundamental because the solution zero structure then achieves its limiting constraint. When $Z_\Omega(M(z)) \approx Z_\Omega$, we refer to $M(z)$ as an essential solution to the model matching equation. Likewise, when $Z_\Gamma(P(z) \approx Z_\Gamma$, we call $P(z)$ an essential solution. Necessary and sufficient conditions are given for the existence of essential solutions. A number of other types of related behavior occur; some of these are discussed in the next section. It should be noted here that the situation for essential solutions in regard to zeros differs concretely from the corresponding scenario for poles. It has been shown by Conte, Perdon, and Wyman [7] that essential solutions for poles always exist in the model matching problem. We shall see momentarily, however, that for zeros they need not exist.

Refer once again to Fig. 4.1, which induces the $k[z]$-linear map $\beta_\Omega(z): Z_\Omega(M(z)) \rightarrow Z_\Omega$. The kernel of this map is given by

$$\ker \beta_\Omega(z) = \{U[z] \cap P^{-1}(PU[z] \cap TR[z])\}/\{U[z] \cap MR[z]\} \tag{6.1a}$$

$$= \{U[z] \cap (MR[z] + \ker P(z))\}/\{U[z] \cap MR[z]\}. \tag{6.1b}$$

A solution $M(z)$ is essential, therefore, when

$$U[z] \cap (MR[z] + \ker P(z)) = U[z] \cap MR[z]. \tag{6.2}$$

On the other hand, from Fig. 4.2, we may determine that the image of $\beta_\Gamma(z): Z_\Gamma \rightarrow Z_\Gamma(P(z))$ is

$$\operatorname{im} \beta_\Gamma(z) = \{M(T^{-1}Y[z]) + P^{-1}Y[z] \cap U[z]\}/\{P^{-1}Y[z] \cap U[z]\}, \tag{6.3a}$$

so that the cokernel is given by isomorphism as

$$\operatorname{coker} \beta_\Gamma(z) \approx P^{-1}Y[z]/\{M(T^{-1}Y[z]) + P^{-1}Y[z] \cap U[z]\}. \tag{6.3b}$$

In this context, then, a solution $P(z)$ is essential when

$$P^{-1}Y[z] = M(T^{-1}Y[z]) + P^{-1}Y[z] \cap U[z]. \tag{6.4}$$

With these preliminaries, we are ready to give the basic result on equalities (6.2) and (6.4).

THEOREM 6. *Let $T(z): R(z) \rightarrow Y(z)$ be a $k(z)$-linear map. If $P(z): U(z) \rightarrow Y(z)$ is a $k(z)$-linear map whose image contains that of $T(z)$, then there exists a $k(z)$-linear map $M(z): R(z) \rightarrow U(z)$ which satisfies the equation $T(z) = P(z)M(z)$ and which has an Ω-zero module*

$$Z_\Omega(M(z)) \approx Z_\Omega \tag{6.5a}$$

if and only if

$$\operatorname{rank} P(z) - \operatorname{rank} T(z) \geqq \dim U - \dim R. \tag{6.5b}$$

If $M(z):R(z)\to U(z)$ *is a* $k(z)$*-linear map whose kernel is contained in that of* $T(z)$*, then there exists a* $k(z)$*-linear map* $P(z):U(z)\to Y(z)$ *which satisfies the equation* $T(z)=P(z)M(z)$ *and which has a* Γ*-zero module*

(6.6a) $$Z_\Gamma(P(z))\approx Z_\Gamma$$

if and only if

(6.6b) $$\operatorname{rank} M(z)-\operatorname{rank} T(z)\geqq \dim U-\dim Y.$$

Proof. For the first part of the theorem, we refer to condition (6.2). If this condition is true, it is necessary that

(6.7) $$U[z]\cap MR[z]\supset U[z]\cap \ker P(z).$$

But the only elements of $MR[z]$ which can possibly be in $\ker P(z)$ are those arising from $\ker T(z)$. Accordingly, (6.7) holds when and only when

(6.8) $$U[z]\cap M(\ker T(z)\cap R[z])\supset U[z]\cap \ker P(z).$$

When considered as a free module, the right member of (6.8) satisfies

(6.9) $$\operatorname{rank}\{U[z]\cap \ker P(z)\}=\dim U-\operatorname{rank} P(z),$$

while the left member depends upon the constraint

(6.10) $$\operatorname{rank}\{\ker T(z)\cap R[z]\}=\dim R-\operatorname{rank} T(z).$$

Together, the preceding three equations imply that (6.5b) must be satisfied. Suppose next that condition (6.5b) is satisfied. We will demonstrate the existence of an $M(z)$ such that (6.2) holds. Observe that $\ker T(z)$ and $\ker P(z)$ are divisible modules, so that

(6.11) $$R(z)=\tilde{R}(z)\oplus \ker T(z),\qquad U(z)=\tilde{U}(z)\oplus \ker P(z),$$

for suitable $k[z]$-linear isomorphisms

(6.12) $$\tilde{R}(z)\approx R(z)/\ker T(z),\qquad \tilde{U}(z)\approx U(z)/\ker P(z).$$

Because $M(\ker T(z))\subset \ker P(z)$, each $M(z)$ satisfying the equation $T(z)=P(z)M(z)$ induces a unique, $k(z)$-linear monomorphism $\bar{M}(z):R(z)/\ker T(z)\to U(z)/\ker P(z)$. The fact that $\bar{M}(z)$ is unique can also be seen as a consequence of the feature that the map induced by $P(z)$ on $U(z)/\ker P(z)$ is monic. If we suppress explicit use of the isomorphisms (6.12), we can think of $\bar{M}(z)$ as a map on $\tilde{R}(z)$ into $\tilde{U}(z)$. Next, we define the action of $M(z)$ on $\ker T(z)$. Let $\hat{M}(z):\ker T(z)\cap R[z]\to \ker P(z)\cap U[z]$ be any $k[z]$-linear epimorphism. $\hat{M}(z)$ induces an epic, $k(z)$-linear map on $\ker T(z)$ onto $\ker P(z)$; and we employ the same notation for both. Together, $\bar{M}(z)$ and $\hat{M}(z)$ can be used to define a map

(6.13) $$\hat{M}(z)\oplus \bar{M}(z):R(z)\to U(z)$$

with the understandings (6.11). For (6.13), we have

$$(\hat{M}(z)\oplus \bar{M}(z))(\ker T(z)\cap R[z]\oplus \tilde{R}(z)\cap R[z])$$

(6.14a) $$=\hat{M}(\ker T(z)\cap R[z])\oplus \bar{M}(\tilde{R}(z)\cap R[z])$$

(6.14b) $$=\ker P(z)\cap U[z]\oplus \bar{M}(\tilde{R}(z)\cap R[z]).$$

Thus, we have

$$U[z]\cap (MR[z]+\ker P(z))$$

(6.14c) $$=U[z]\cap\{(\ker P(z)\cap U[z]+\ker P(z))\oplus \bar{M}(\tilde{R}(z)\cap R[z])\}$$

(6.14d) $$=\ker P(z)\cap U[z]\oplus U[z]\cap \bar{M}(\tilde{R}(z)\cap R[z])$$

(6.14e) $$=U[z]\cap MR[z],$$

as desired. Note that, for these steps to be valid, the formation of (6.11) and (6.12) must satisfy

(6.14f) $$R[z]=\tilde{R}(z)\cap R[z]\oplus\ker T(z)\cap R[z],$$

(6.14g) $$U[z]=\tilde{U}(z)\cap U[z]\oplus\ker P(z)\cap U[z].$$

This is always possible. For the second part of the theorem, we make use of the condition expressed by (6.4). Because im $M(z)$ and im $T(z)$ are divisible modules, we can write

(6.15) $$U(z)=\tilde{U}(z)\oplus\operatorname{im} M(z), \qquad Y(z)=\tilde{Y}(z)\oplus\operatorname{im} T(z),$$

for suitable, $k[z]$-linear, isomorphisms

(6.16a) $$\tilde{U}(z)\approx\operatorname{coker} M(z), \qquad \tilde{Y}(z)\approx\operatorname{coker} T(z),$$

satisfying the constraints

(6.16b) $$U[z]=\tilde{U}(z)\cap U[z]\oplus\operatorname{im} M(z)\cap U[z],$$

(6.16c) $$Y[z]=\tilde{Y}(z)\cap Y[z]\oplus\operatorname{im} T(z)\cap Y[z].$$

Restricted to im $M(z)$, $P(z)$ gives an epic, $k(z)$-linear map $\hat{P}(z)$ onto im $T(z)$. Moreover, the action of $\hat{P}(z)$ remains the same for any solution $P(z)$ of the equation $T(z)=P(z)M(z)$. The design freedom for $P(z)$ lies, therefore, in its action on the cokernel of $M(z)$, represented by $\tilde{U}(z)$. Observe that ker $T(z)$ is a divisible submodule of $T^{-1}Y[z]$. Accordingly,

(6.17a) $$T^{-1}Y[z]=\ker T(z)\oplus\hat{R}(z),$$

where $\hat{R}(z)$ satisfies

(6.17b) $$\hat{R}(z)\approx T^{-1}Y[z]/\ker T(z),$$

as a $k[z]$-module. Though $\hat{R}(z)$ is finitely generated, ker $T(z)$ is not; and so the only term in the right member of (6.4) which is not finitely generated is $M(\ker T(z))$. We have, then, the preliminary necessity of

(6.18) $$M(\ker T(z))\supset\ker P(z),$$

which implies that the $k[z]$-linear map $\bar{P}(z)$: coker $M(z)\to$ coker $T(z)$, induced by $P(z)$, is a monomorphism. But this means that

(6.19) $$\dim(\operatorname{coker} T(z))\geqq\dim(\operatorname{coker} M(z)),$$

from which we have the condition

(6.20) $$\dim Y-\operatorname{rank} T(z)\geqq\dim U-\operatorname{rank} M(z),$$

which gives (6.6b). Under (6.6b), we can construct an essential solution. Let $\bar{P}(z)$ be any $k(z)$-linear monomorphism on $\tilde{U}(z)$ into $\tilde{Y}(z)$ with the property that

(6.21) $$\bar{P}^{-1}(\tilde{Y}(z)\cap Y[z])\subset\tilde{U}(z)\cap U[z].$$

Together, $\hat{P}(z)$ and $\bar{P}(z)$ give the map

(6.22) $$\hat{P}(z)\oplus\bar{P}(z): U(z)\to Y(z),$$

for which

$$P^{-1}Y[z] = P^{-1}\{\tilde{Y}(z)\cap Y[z]\oplus \text{im } T(z)\cap Y[z]\} \tag{6.23a}$$

$$= \bar{P}^{-1}(\tilde{Y}(z)\cap Y[z])\oplus \hat{P}^{-1}(\text{im } T(z)\cap Y[z]) \tag{6.23b}$$

$$\subset P^{-1}Y[z]\cap U[z]+\text{im } M(z)\cap P^{-1}Y[z] \tag{6.23c}$$

$$= P^{-1}Y[z]\cap U[z]+M(T^{-1}Y[z]), \tag{6.23d}$$

as required.

Discussion of Theorem 6. The conditions of the theorem, of course, are not always satisfied. This means that essential solutions for model matching problems need not always exist with respect to zeros. As explained in the prologue to this section, such a situation differs from that in the case of poles. We might ask if, when attention is fixed upon finitely generated, torsion zeros, a result more like that for poles could be obtained. We will see in the next section that the answer to this question must be given as a negative. When $P(z)$ is given, the construction of the theorem has to do with designing $\hat{M}(z)$ in such a way that

$$Z_\Omega(\hat{M}(z)) = 0. \tag{6.24a}$$

In like fashion, when $M(z)$ is given, the goal is a design of $\bar{P}(z)$ so that

$$Z_\Gamma(\bar{P}(z)) = 0. \tag{6.24b}$$

In fact, for $M(z)$ of the form (6.13), we have

$$\ker \beta_\Omega(z) \approx Z_\Omega(\hat{M}(z)), \tag{6.25a}$$

while for $P(z)$ of the form (6.22) we find

$$\text{coker } \beta_\Gamma(z) \approx Z_\Gamma(\bar{P}(z)). \tag{6.25b}$$

For instance, (6.25a) can be seen from a construction

$$\{U[z]\cap(MR[z]+\ker P(z))\}/\{U[z]\cap MR[z]\}$$

$$= \{U[z]\cap(MR[z]+\ker P(z))+U[z]\cap MR[z]\}/\{U[z]\cap MR[z]\} \tag{6.26a}$$

$$= \{U[z]\cap(\hat{M}(\ker T(z)\cap R[z])+\ker P(z))+U[z]\cap MR[z]\}$$

$$/\{\text{U}[z]\cap MR[z]\} \tag{6.26b}$$

$$= \{U[z]\cap \ker P(z)+U[z]\cap MR[z]\}/\{U[z]\cap MR[z]\} \tag{6.26c}$$

$$\approx \{U[z]\cap \ker P(z)\}/\{U[z]\cap \ker P(z)\cap \hat{M}(\ker T(z)\cap R[z])\}. \tag{6.26d}$$

More generally, whenever $M(z)$ is a solution to the equation $T(z) = P(z)M(z)$, whether or not it takes the form (6.13), we have that $Z_\Omega(M(z)|\ker T(z))$ is a submodule of $Z_\Omega(M(z))$. In the same way, whenever $P(z)$ is a solution and $\bar{P}(z)$ is its induced map on coker $M(z)$, then $Z_\Gamma(\bar{P}(z))$ is a factor module of $Z_\Gamma(P(z))$, whether or not $P(z)$ takes the form (6.22). These facts may be deduced from Figs. 6.1 and 6.2, following. In the first, Fig. 6.1, both rows and columns are natural inclusions. In the second, Fig. 6.2, row one is an inclusion, while the columns are natural projections; row two is a monomorphism of $k[z]$-modules.

$$\begin{tikzcd}
U[z]\cap MR[z]\cap \ker P(z) \arrow[r] \arrow[d] & U[z]\cap \ker P(z) \arrow[d] \\
U[z]\cap MR[z] \arrow[r] & U[z]
\end{tikzcd}$$

FIG. 6.1

$$\begin{tikzcd}
P^{-1}Y[z]\cap U[z] \arrow[r] \arrow[d] & P^{-1}Y[z] \arrow[d] \\
\frac{P^{-1}Y[z]\cap U[z]}{M(T^{-1}Y[z])\cap U[z]} \arrow[r] & \frac{P^{-1}Y[z]}{M(T^{-1}Y[z])}
\end{tikzcd}$$

FIG. 6.2

Remark. In case ker $T(z)$ is equal to zero, the rank of $T(z)$ must equal the dimension of R. Thus (6.5b) leads to $P(z)$ having zero kernel as well. If coker $T(z)$ equals zero, then (6.6b) gives that $M(z)$ must have zero cokernel as well. Note that in these cases the solutions, when they exist, are unique. Moreover, the unique solution $M(z)$ in (6.5a) has no Γ-zeros; and the unique solution $P(z)$ in (6.6a) has no Ω-zeros. In fact, much more is true. See the remarks after Theorem 7, where it is shown in an alternative context that the vanishing of ker $P(z)$ or coker $M(z)$ is always sufficient to satisfy (6.5b) or (6.6b), respectively. This is true whether or not we take special cases of $T(z)$.

The existence of solutions to the model matching equation $T(z)=P(z)M(z)$ is controlled by the conditions im $T(z)\subset$ im $P(z)$ and ker $M(z)\subset$ ker $T(z)$, respectively. In practice, the marginal case of these conditions is of great interest.

COROLLARY 1. *Under the condition* im $T(z)=$ im $P(z)$, *there exists an essential solution $M(z)$ if and only if*

$$\dim R \geqq \dim U. \tag{6.27a}$$

Under the condition ker $M(z)=$ ker $T(z)$, *there exists an essential solution $P(z)$ if and only if*

$$\dim Y \geqq \dim U. \tag{6.27b}$$

Another case of special interest occurs when the domains of $T(z)$ and $P(z)$, or the codomains of $T(z)$ and $M(z)$, have equal dimensions. In feedback systems, for instance, this situation has strong physical meaning.

COROLLARY 2. *Let the dimensions of R and U be equal. Then, whenever a solution $M(z)$ exists, there exists an essential solution. Alternatively, let the dimensions of Y and U be equal. Then, whenever a solution $P(z)$ exists, there exists an essential solution.*

In the case of feedback control, $M(z)$ stands for an open loop pre-compensator which is equivalent at specified operating conditions to a given closed-loop strategy. The concept of an essential solution carries a pleasing connotation in this case. For a solution $M(z)$ which is not essential presents a certain deficiency in its image; and this in turn points out an inadequacy in manipulating the control inputs to the plant $P(z)$. One way to address this difficulty is to increase the dimension of R, which corresponds to adding additional reference inputs; however, this method is subject to a suitable re-interpretation of $T(z)$. Another way, which is more traditional, is to reduce the dimension of U, by deleting dependent columns of $P(z)$, which corresponds to using fewer controls. Before the concept of extended zeros, the effects of plant input selection have been seen only in terms of their influence upon finitely generated, torsion zeros. In view of the preceding results on matching modules of extended zeros, however, this traditional approach may need to be reconsidered.

The matching zero modules, and the extended zero modules of solutions to model matching equations, treat finitely generated, torsion zeros together with divisible or free zeros, as the case may be. It is sometimes desirable to have in hand an idea of what is happening in regard to these individual types of zeros. Not surprisingly, it turns out that these behaviors are interrelated, by a snake mapping; and this is the subject of the next section.

7. The fundamental diagrams. Sections 3-6 have placed a focus upon the extended zero modules of Γ-type and of Ω-type and their role in model matching problems. In this section, our goal is to relate these results to solution zeros of finitely generated, torsion type, as described in § 2, and to divisible or free zeros, as explained in § 3. The main ideas center upon a pair of commutative diagrams.

THEOREM 7. *Suppose that $T(z): R(z) \to Y(z)$ is a $k(z)$-linear map. If $P(z): U(z) \to Y(z)$ is a $k(z)$-linear map whose image contains that of $T(z)$, and if $M(z): R(z) \to U(z)$ is a $k(z)$-linear map satisfying $T(z) = P(z)M(z)$, then there exist $k[z]$-linear maps $\hat{\beta}_\Omega(z): Z(M(z)) \to Z'_\Omega$ and $\bar{\beta}_\Omega(z): \Omega(M(z)) \to \Omega$, the latter an epimorphism, such that the diagram Fig.* 7.1 *commutes and has rows which are short exact sequences. If $M(z): R(z) \to U(z)$ is a $k(z)$-linear map whose kernel is included in that of $T(z)$, and if $P(z): U(z) \to Y(z)$ is a $k(z)$-linear map satisfying $T(z) = P(z)M(z)$, then there exist $k[z]$-linear maps $\hat{\beta}_\Gamma(z): \Gamma \to \Gamma(P(z))$ and $\bar{\beta}_\Gamma(z): Z'_\Gamma \to Z(P(z))$, the former a monomorphism, such that the diagram Fig.* 7.2 *commutes and has rows which are short exact sequences.*

Proof. Consider Fig. 7.1, and note that column two is the map (4.21a). Row one is (3.9), constructed as in § 3.2; and row two is just (4.17). Because $\beta_\Omega Z(M(z)) \subset Z'_\Omega$, the map $\hat{\beta}_\Omega(z)$ exists uniquely; and, consequently, there is a unique $\bar{\beta}_\Omega(z)$ which completes the diagram. A construction for Fig. 7.2 can be accomplished in like manner, with the aid of (4.21b), (3.4), and (4.19).

Theorem 7 leads to a number of quite detailed consequences. The first of these is a pair of technical corollaries.

COROLLARY 3. *Under the assumptions of Theorem* 7, *there exist $k[z]$-linear maps such that the following pair of sequences is exact:*

(7.1) $$0 \to \ker \hat{\beta}_\Omega(z) \to \ker \beta_\Omega(z) \to \ker \bar{\beta}_\Omega(z) \to \operatorname{coker} \hat{\beta}_\Omega(z) \to 0,$$

(7.2) $$0 \to \ker \bar{\beta}_\Gamma(z) \to \operatorname{coker} \hat{\beta}_\Gamma(z) \to \operatorname{coker} \beta_\Gamma(z) \to \operatorname{coker} \bar{\beta}_\Gamma(z) \to 0.$$

$$\begin{tikzcd}
0 \arrow[r] & Z(M(z)) \arrow[r] \arrow[d, "\hat{\beta}_\Omega(z)"'] & Z_\Omega(M(z)) \arrow[r] \arrow[d, "\beta_\Omega(z)"'] & \Omega(M(z)) \arrow[r] \arrow[d, "\bar{\beta}_\Omega(z)"'] & 0 \\
0 \arrow[r] & Z'_\Omega \arrow[r] & Z_\Omega \arrow[r] \arrow[d] & \Omega \arrow[r] \arrow[d] & 0 \\
 & & 0 & 0 &
\end{tikzcd}$$

FIG. 7.1

$$\begin{tikzcd}
 & 0 \arrow[d] & 0 \arrow[d] & & \\
0 \arrow[r] & \Gamma \arrow[r] \arrow[d, "\hat{\beta}_\Gamma(z)"'] & Z_\Gamma \arrow[r] \arrow[d, "\beta_\Gamma(z)"'] & Z'_\Gamma \arrow[r] \arrow[d, "\bar{\beta}_\Gamma(z)"'] & 0 \\
0 \arrow[r] & \Gamma(P(z)) \arrow[r] & Z_\Gamma(P(z)) \arrow[r] & Z(P(z)) \arrow[r] & 0
\end{tikzcd}$$

FIG. 7.2

Proof. Sequences (7.1) and (7.2) are consequences of the Snake lemma, applied to the pair of diagrams, Figs. 7.1 and 7.2.

COROLLARY 4. *Under the assumptions of Theorem* 7, *the modules in sequences* (7.1) *and* (7.2) *are given, up to* $k[z]$*-linear isomorphism, by*

$$\ker \hat{\beta}_\Omega(z) = \{U[z] \cap (MR[z] + \ker P(z) \cap MR(z))\}/\{U[z] \cap MR[z]\}, \tag{7.3a}$$

$$\ker \beta_\Omega(z) = \{U[z] \cap (MR[z] + \ker P(z))\}/\{U[z] \cap MR[z]\}, \tag{7.3b}$$

$$\ker \bar{\beta}_\Omega(z) = \{U[z] \cap (MR(z) + \ker P(z))\}/\{U[z] \cap MR(z)\}, \tag{7.3c}$$

$$\operatorname{coker} \hat{\beta}_\Omega(z) = \{PU[z] \cap TR(z)\}/\{P[U[z] \cap MR(z)] + PU[z] \cap TR[z]\}, \tag{7.3d}$$

$$\begin{aligned}\ker \bar{\beta}_\Gamma(z) = M^{-1}\{&\ker P(z) + P^{-1}Y[z] \cap U[z]\}\\ &/\{\ker T(z) + T^{-1}Y[z] \cap M^{-1}U[z]\},\end{aligned} \tag{7.4a}$$

$$\operatorname{coker} \hat{\beta}_\Gamma(z) = \ker P(z)/\{M(\ker T(z)) + \ker P(z) \cap U[z]\}, \tag{7.4b}$$

$$\operatorname{coker} \beta_\Gamma(z) = P^{-1}Y[z]/\{M(T^{-1}Y[z]) + P^{-1}Y[z] \cap U[z]\}, \tag{7.4c}$$

$$\operatorname{coker} \bar{\beta}_\Gamma(z) = P^{-1}Y[z]/\{M(T^{-1}Y[z]) + \ker P(z) + P^{-1}Y[z] \cap U[z]\}. \tag{7.4d}$$

Proof. We have already discussed (7.3b) and (7.4c) in § 6. The remaining forms follow from standard diagram chases applied to (4.21) and the row maps in (3.4), (3.9), (4.17), and (4.19). For brevity, we omit the details.

Remark. Observe that, when $\ker P(z)$ is zero, (7.3a), (7.3b), and (7.3c) vanish. From Fig. 7.1, we then have that (7.3d) vanishes. In this case, $Z_\Omega(M(z))$ is isomorphic as a $k[z]$-module to Z_Ω. If a solution $M(z)$ exists, therefore, $Z_\Omega(M(z))$ can fail to achieve its bound only if there is the possibility of more than one solution. Similar statements can be made for $\operatorname{coker} M(z)$, in reference to (7.4a), (7.4b), (7.4c), and (7.4d), relative to Fig. 7.2.

THEOREM 8. *Suppose that* $T(z) = P(z)M(z)$ *is an equation of* $k(z)$*-linear maps. Then each of the following three statements implies the others*:

$$\Omega(M(z)) \approx \Omega, \tag{7.5a}$$

$$\Gamma(P(z)) \approx \Gamma, \tag{7.5b}$$

$$M(\ker T(z)) = \ker P(z). \tag{7.5c}$$

In the case where (7.5) *holds, then* Z'_Ω *is a factor module of* $Z(M(z))$ *and* Z'_Γ *is a submodule of* $Z(P(z))$.

Proof. Condition (7.5a) makes $\ker \bar{\beta}_\Omega(z)$ vanish, which by (7.1) makes $\operatorname{coker} \hat{\beta}_\Omega(z)$ vanish, which means that Z'_Ω is a factor module of $Z(M(z))$. A like argument establishes that Z'_Γ is a submodule of $Z(P(z))$. From (7.5c), we find that (7.4b) is zero, which implies (7.5b); again, (7.5c) gives (7.3c) equal to zero, which leads to (7.5a). Now consider (7.5b), which implies that (7.4b) must vanish. If (7.5c) does not hold, then (7.4b) cannot vanish because $\ker P(z) \cap U[z]$ is finitely generated, while $\ker P(z)$ is not. Thus (7.4b) equal to zero must give (7.5c). Finally, we look at (7.5a), for which (7.3c) must vanish. A necessary condition for (7.3c) equal to zero is that

$$U[z] \cap MR(z) \supset U[z] \cap \ker P(z), \tag{7.6a}$$

which holds only if

$$U[z] \cap M(\ker T(z)) \supset U[z] \cap \ker P(z), \tag{7.6b}$$

which requires

$$\dim M(\ker T(z)) \geqq \dim \ker P(z); \tag{7.6c}$$

and (7.5c) follows.

Discussion of Theorem 8. The coincidence of the three statements (7.5) suggests the possibility that a corresponding observation may relate (6.5a) to (6.6a) in Theorem 6. However, such is not the case. For a given $P(z)$, when (6.5b) is satisfied, we may construct an $M(z)$ that fulfills (6.5a). Now, for this $M(z)$, it is quite possible that the original $P(z)$ fails to satisfy (6.6a). To demonstrate this behavior, we present an example.

Example 1.

(7.7)
$$T(z)=\begin{bmatrix}1&0\\0&0\\0&0\end{bmatrix},\quad P(z)=\begin{bmatrix}1&0&0\\0&0&0\\0&z&0\end{bmatrix},\quad M(z)=\begin{bmatrix}1&0\\0&0\\0&1\end{bmatrix}.$$

In this example, $M(z)$ is an essential solution; but $P(z)$ is not. On the other hand, conditions (7.5) hold.

When conditions (7.5) are satisfied, we have that both (6.5b) and (6.6b) hold simultaneously. If $P(z)$ is given, with (6.5b) true, then we can design an $M(z)$ so that (7.5) occurs. One way to carry out such a construction is to proceed according to the discussion following (6.11) and (6.12); for this situation, however, $\hat{M}(z)$ can be any epic, $k(z)$-linear map from ker $T(z)$ onto ker $P(z)$. Again, if $M(z)$ is given, with (6.6b) true, we can design a $P(z)$ so that (7.5) takes place. The design may be executed along the lines established in the proof of Theorem 6, except that we do not need the condition (6.21). In a manner similar to that of (6.24), the basic constructive conditions for these solutions are

(7.8a)
$$\Omega(\hat{M}(z))=0,$$

(7.8b)
$$\Gamma(\bar{P}(z))=0.$$

These follow from the $k[z]$-linear isomorphisms

(7.8c)
$$\ker\bar{\beta}_{\Omega}(z)\approx\Omega(\hat{M}(z)),$$

(7.8d)
$$\operatorname{coker}\hat{\beta}_{\Gamma}(z)\approx\Gamma(\bar{P}(z)),$$

for an $M(z)$ of form (6.13) or a $P(z)$ of form (6.22). In fact, for more general types of solutions $M(z)$ or $P(z)$, and with the aid of diagrams analogous to Figs. 6.1 and 6.2, we can establish that $\Omega(M(z)|\ker T(z))$ is a submodule of $\Omega(M(z))$, while $\Gamma(\bar{P}(z))$ is a factor module of $\Gamma(P(z))$. Here we denote by $\bar{P}(z)$ the induced map on coker $M(z)$.

Next we turn to a study of the finitely generated, torsion zeros in solutions to the model matching equation. In general, these solutions have zeros arising from the finitely generated, torsion parts of Z_{Γ} or Z_{Ω}. They may also have zeros which occur through the design of the induced action of $P(z)$ on coker $M(z)$ or of the restricted action of $M(z)$ on ker $T(z)$. Additional effects can also occur.

$$\begin{tikzcd}
U[z]\cap M(\ker T(z)\cap R[z]) \arrow[r] \arrow[d] & U[z]\cap M(\ker T(z)) \arrow[d] \\
U[z]\cap MR[z] \arrow[r] & U[z]\cap(MR[z]+M(\ker T(z)))
\end{tikzcd}$$

FIG. 7.3

We begin by directing our attention to the commutative diagram of Fig. 7.3, wherein the rows and columns are natural inclusions. This diagram induces a $k[z]$-linear map on $Z(M(z)|\ker T(z))$ into $\ker\hat{\beta}_{\Omega}(z)$. The kernel of this map is given by

(7.9)
$$\{U[z]\cap MR[z]\cap M(\ker T(z))\}/\{U[z]\cap M(\ker T(z)\cap R[z])\}.$$

But this will vanish if

$$M(\ker T(z) \cap R[z]) = M \ker T(z) \cap MR[z], \tag{7.10a}$$

which occurs when

$$(\ker T(z) + R[z]) \cap \ker M(z) = \ker T(z) \cap \ker M(z) + R[z] \cap \ker M(z). \tag{7.10b}$$

Consider an element in the left member of (7.10b); and represent it by

$$r_t(z) + r(z) = r_m(z). \tag{7.10c}$$

Applying $T(z)$ to (7.10c) shows that $r(z) \in \ker T(z) \cap R[z]$; and so both members of (7.10b) are equal to $\ker M(z)$. The induced map is therefore monic; and $Z(M(z)|\ker T(z))$ is a submodule of $\ker \hat{\beta}_\Omega(z)$, which is itself a submodule of $Z(M(z))$. In this way, we can use $M(z)|\ker T(z)$ to design specified zero submodules for solutions $M(z)$. The cokernel of the induced map is

$$\{U[z] \cap (MR[z] + M(\ker T(z)))\}/\{U[z] \cap M(\ker T(z)) + U[z] \cap MR[z]\}. \tag{7.11}$$

If $M(z)$ may be regarded as a direct sum map on $\ker T(z) \oplus \tilde{R}(z)$ into $\ker P(z) \oplus \tilde{U}(z)$, for $\tilde{R}(z)$ and $\tilde{U}(z)$ defined as in (6.11) and (6.12), and satisfying the conditions (6.14), then

$$MR[z] + M(\ker T(z)) = M[\tilde{R}(z) \cap R[z]] + M(\ker T(z)), \tag{7.12}$$

and (7.11) is zero, with $\ker \hat{\beta}_\Omega(z) \approx Z(\hat{M}(z))$. However, if $M(z)$ is otherwise, (7.11) may not disappear; and the solution may display additional zeros. To illustrate, we give an example.

Example 2.

$$T(z) = [1 \;\; 0], \quad P(z) = [1 \;\; 0], \quad M(z) = \begin{bmatrix} 1 & 0 \\ 1/(z+1) & 1 \end{bmatrix}. \tag{7.13}$$

Here, we have $Z'_\Omega = 0$, $Z(M(z)|\ker T(z)) = 0$, and $Z(M(z)) \approx k[z]/(z+1)k[z]$. The solution zero arises entirely from (7.11).

Another key observation can be made. If im $M(z)$ includes $\ker P(z)$, then from Theorem 8 we know that

$$Z(M(z))/\ker \hat{\beta}_\Omega(z) \approx Z'_\Omega, \tag{7.14}$$

which means that the torsion part of Z_Ω is necessarily present in $Z(M(z))$ as a factor module.

Turning, then, to the possibility of removing some part of Z'_Ω from our solution, we see that decreasing the dimension of $M(\ker T(z))$ is to our advantage. The best we can achieve in this regard is to choose $M(z)|\ker T(z)$ equal to zero. In this case, $\ker \hat{\beta}_\Omega(z)$ vanishes, and $Z(M(z))$ is a submodule of Z'_Ω. If the number of nonunit invariant factors of Z'_Ω is no greater than the dimension of $\ker P(z)$, then a solution $M(z)$ with $Z(M(z)) = 0$ is possible. For instance, if the zeros of Z'_Ω are all distinct, then $Z(M(z)) = 0$ is possible whenever $M(z)$ is not unique. Thus there is a tradeoff between inserting into the solution zeros which are distinct from those in Z'_Ω and removing from the solution zeros which are part of Z'_Ω. We summarize in the next theorem.

THEOREM 9. *Suppose that $T(z): R(z) \to Y(z)$ is a $k(z)$-linear map. If $P(z): U(z) \to Y(z)$ is a $k(z)$-linear map whose image contains that of $T(z)$, then $Z(M(z)|\ker T(z))$ is a submodule of $Z(M(z))$ whenever a $k(z)$-linear map $M(z): R(z) \to U(z)$ satisfies the equation $T(z) = P(z)M(z)$. A maximum of*

$$\min(\dim \ker P(z), \dim \ker T(z)) \tag{7.15}$$

invariant factors of $Z(M(z))$ can be specified in this way. Alternatively, up to

$$\dim \ker P(z) - \operatorname{rank} M(z)|\ker T(z) \tag{7.16}$$

nonunit invariant factors of Z'_Ω can be made into units in $Z(M(z))$ by design of $M(z)$ off the kernel of $T(z)$. If $M(z): R(z) \to U(z)$ is a $k(z)$-linear map whose kernel is included in that of $T(z)$, and if $P(z)$ is a $k(z)$-linear solution to the equation $T(z) = P(z)M(z)$, then $Z(\bar{P}(z))$ is a factor module of $Z(P(z))$, where $\bar{P}(z)$ is the map induced on coker $M(z)$. *A maximum of*

$$\min(\dim \operatorname{coker} T(z), \dim \operatorname{coker} M(z)) \tag{7.17}$$

invariant factors of $Z(P(z))$ can be specified in this way. Alternatively, up to

$$\dim \operatorname{coker} M(z) - \operatorname{rank} \bar{P}(z) \tag{7.18}$$

nonunit invariant factors of Z'_Γ can be made into units in $Z(P(z))$ by design of $P(z)$ off the cokernel of $M(z)$.

Proof. The argument for the first part of the theorem has been presented in the prologue. In the interest of space, we omit the second half.

Discussion of Theorem 9. It should be noted that the statements in this result have a number of interesting modifications. For example, with reference to (7.16), we do not have to make the invariant factor of Z'_Ω into a unit; instead, we could make a more general adjustment to it. Again, in regard to (7.18), adjustment of invariant factors could replace their reduction to units.

From Theorem 9, we see that Z'_Ω and Z'_Γ play an important role in the character of $Z(M(z))$ and $Z(P(z))$, respectively. The next section inquires further into the nature of these modules.

8. Description of Z'_Ω and Z'_Γ. In § 4, we have shown that the matching zero module Z_Ω, defined in (4.6), appears as a factor module in the zero module $Z(M(z))$ of every solution $M(z)$ to the model matching equation. Moreover, it was shown that Z_Ω has a torsion submodule Z'_Ω as given in (4.9). From these two results, it is natural to inquire about the nature of Z'_Ω.

We begin with the commutative diagram of Fig. 8.1. In this diagram, both rows and columns are natural inclusions. Notice, in the figure, that the second column, when extended to a short exact sequence, has its factor module isomorphic to Z'_Ω. Observe also that the first row extends into a short exact sequence with factor module $P(T(z))$. From Fig. 8.1, we find a monic, $k[z]$-linear map from $P(T(z))$ into the module

$$_T X_P = \frac{PU[z] \cap TR(z) + TR[z]}{Y[z] \cap \{PU[z] \cap TR(z) + TR[z]\}}. \tag{8.1}$$

The cokernel of this map is denoted by X'_z. It is possible to relate the module (8.1) to the pole module of the composite map $[T(z) \ P(z)]: R(z) \oplus U(z) \to Y(z)$. The basic idea is shown in the commutative diagram of Fig. 8.2, where rows and columns are

$$\begin{tikzcd}
Y[z] \cap TR[z] \arrow[r] \arrow[d] & TR[z] \arrow[d] \\
Y[z] \cap \{P\dot{U}[z] \cap TR(z) + TR[z]\} \arrow[r] & PU[z] \cap TR(z) + TR[z]
\end{tikzcd}$$

FIG. 8.1

$$\begin{tikzcd}
Y[z] \cap \{PU[z] \cap TR(z) + TR[z]\} \arrow[r] \arrow[d] & PU[z] \cap TR(z) + TR[z] \arrow[d] \\
Y[z] \cap \{PU[z] + TR[z]\} \arrow[r] & PU[z] + TR[z]
\end{tikzcd}$$

FIG. 8.2

again natural inclusions. From this diagram, we induce a monic, $k[z]$-linear map on ${}_TX_P$ into $P([T(z)\ \ P(z)])$, with cokernel isomorphic to

$$X_k = \frac{PU[z] + TR[z]}{PU[z] \cap TR(z) + TR[z] + Y[z] \cap \{PU[z] + TR[z]\}}. \tag{8.2}$$

Returning now to Fig. 8.1, we define the factor module induced by column one to be Z'_z, and point out that it is a submodule of $Z(T(z))$ by the inclusion

$$PU[z] \cap TR(z) + TR[z] \subset TR(z). \tag{8.3}$$

A natural inclusion of Z'_z into $Z(T(z))$ induces a factor module isomorphic to

$${}_TZ_P = \frac{Y[z] \cap TR(z)}{Y[z] \cap \{PU[z] \cap TR(z) + TR[z]\}}. \tag{8.4}$$

Module (8.4) can be related to the zero module $Z([T(z)\ \ P(z)])$. Refer to the commutative diagram of Fig. 8.3 of natural inclusions. This diagram induces a monic, $k[z]$-linear map on ${}_TZ_P$ into $Z([T(z)\ \ P(z)])$, with cokernel isomorphic to

$$Z_k = \frac{Y[z] \cap PU(z)}{Y[z] \cap TR(z) + Y[z] \cap (TR[z] + PU[z])}. \tag{8.5}$$

We summarize these results.

THEOREM 10A. *Let* im $T(z) \subset$ im $P(z)$, *and let* Z'_Ω *be the torsion submodule of the matching zero module* Z_Ω. *Then there exist* $k[z]$*-modules* ${}_TZ_P$, Z_k, ${}_TX_P$, X_k, Z'_z, *and* X'_z, *together with appropriate* $k[z]$*-linear maps, such that the following five short sequences are exact*:

$$0 \to {}_TX_P \to P([T(z)\ \ P(z)]) \to X_k \to 0, \tag{8.6a}$$

$$0 \to P(T(z)) \to {}_TX_P \to X'_z \to 0, \tag{8.6b}$$

$$0 \to {}_TZ_P \to Z([T(z)\ \ P(z)]) \to Z_k \to 0, \tag{8.6c}$$

$$0 \to Z'_z \to Z(T(z)) \to {}_TZ_P \to 0, \tag{8.6d}$$

$$0 \to Z'_z \to Z'_\Omega \to X'_z \to 0. \tag{8.6e}$$

Proof. Sequences (8.6a)-(8.6d) have been established in the prelude to the theorem. Sequence (8.6e) is a consequence of the requirements on cokernels of columns in Fig. 8.1, including the induced column.

$$\begin{tikzcd}
Y[z] \cap \{PU[z] \cap TR(z) + TR[z]\} \arrow[r] \arrow[d] & Y[z] \cap TR(z) \arrow[d] \\
Y[z] \cap \{TR[z] + PU[z]\} \arrow[r] & Y[z] \cap \{TR(z) + PU(z)\} = Y[z] \cap PU(z)
\end{tikzcd}$$

FIG. 8.3

Discussion of Theorem 10A. The study of Z'_Ω then begins with the finitely generated, torsion poles and zeros of $T(z)$ and of $[T(z)\ P(z)]$. From the submodules ${}_TZ_P$ of $Z([T(z)\ P(z)])$ and ${}_TX_P$ of $P([T(z)\ P(z)])$, together with $Z(T(z))$ and $P(T(z))$, we can find Z'_z and X'_z, the building blocks of Z'_Ω. Note that, if $\operatorname{im} T(z) = \operatorname{im} P(z)$, both Z_k and X_k vanish. In that case, we obtain the sequences

$$0 \to P(T(z)) \to P([T(z)\ P(z)]) \to X'_z \to 0, \tag{8.7a}$$

$$0 \to Z'_z \to Z(T(z)) \to Z([T(z)\ P(z)]) \to 0, \tag{8.7b}$$

$$0 \to Z'_z \to Z'_\Omega \to X'_z \to 0, \tag{8.7c}$$

and this result then reduces to Theorem 4, because Ω is zero. In general, however, the description of Z'_Ω is more complicated than that of Z_Ω, with five sequences instead of three.

Consider next the commutative diagram of Fig. 8.4, where columns and rows are natural inclusions. It is clear that row two extends into a short exact sequence having $Z(T(z))$ as the factor module. The corresponding construction on row one produces a module

$$\frac{T^{-1}Y[z] \cap (\ker T(z) + M^{-1}U[z])}{T^{-1}Y[z] \cap (\ker T(z) + M^{-1}U[z]) \cap (\ker T(z) + R[z])} = Z_M^T, \tag{8.8}$$

which is finitely generated and torsion. Moreover, a brief diagram chase yields the existence of a monic $k[z]$-linear map on Z_M^T into $Z(T(z))$, with cokernel Z''_z. The character of Z_M^T follows from the calculation

$$\frac{T^{-1}Y[z] \cap (\ker T(z) + M^{-1}U[z])}{T^{-1}Y[z] \cap (\ker T(z) + M^{-1}U[z]) \cap (\ker T(z) + R[z])}$$

$$\approx \frac{T^{-1}Y[z] \cap (\ker T(z) + M^{-1}U[z]) + \ker T(z) + R[z]}{\ker T(z) + R[z]} \tag{8.9a}$$

$$= \frac{T^{-1}Y[z] \cap M^{-1}U[z] + \ker T(z) + R[z]}{\ker T(z) + R[z]} \tag{8.9b}$$

$$\approx \frac{T^{-1}Y[z] \cap M^{-1}U[z]}{T^{-1}Y[z] \cap M^{-1}U[z] \cap (\ker T(z) + R[z])}. \tag{8.9c}$$

Then, as a result of the inclusion

$$\begin{aligned} &T^{-1}Y[z] \cap M^{-1}U[z] \cap (\ker M(z) + R[z]) \\ &\qquad \subset T^{-1}Y[z] \cap M^{-1}U[z] \cap (\ker T(z) + R[z]), \end{aligned} \tag{8.10}$$

there exists a $k[z]$-linear epimorphism from

$$Z\left(\begin{bmatrix} T(z) \\ M(z) \end{bmatrix}\right) \tag{8.11}$$

$$\begin{tikzcd}
T^{-1}Y[z] \cap (\ker T(z) + M^{-1}U[z]) \cap (\ker T(z) + R[z]) \arrow[r] \arrow[d] & T^{-1}Y[z] \cap (\ker T(z) + M^{-1}U[z]) \arrow[d] \\
T^{-1}Y[z] \cap (\ker T(z) + R[z]) \arrow[r] & T^{-1}Y[z]
\end{tikzcd}$$

FIG. 8.4

onto Z_M^T, with kernel isomorphic to

$$Z'_k = \frac{(\ker T(z)+R[z]) \cap T^{-1}Y[z] \cap M^{-1}U[z]}{(\ker M(z)+R[z]) \cap T^{-1}Y[z] \cap M^{-1}U[z]}. \tag{8.12}$$

We turn now to column one of Fig. 8.4; and we wish to relate the cokernel to certain pole modules. Note that

$$P(T(z)) = \frac{R[z]}{T^{-1}Y[z] \cap R[z]} \approx \frac{R[z]+\ker T(z)}{T^{-1}Y[z] \cap (R[z]+\ker T(z))}; \tag{8.13}$$

and consider the module

$$\begin{aligned} &\frac{R[z]}{(\ker T(z)+T^{-1}Y[z] \cap M^{-1}U[z]) \cap R[z]} \\ &\approx \frac{\ker T(z)+R[z]}{T^{-1}Y[z] \cap (\ker T(z)+M^{-1}U[z]) \cap (\ker T(z)+R[z])}. \end{aligned} \tag{8.14}$$

By the inclusion of denominators in right members of (8.13) and (8.14), there is an epic $k[z]$-linear map from X_M^T, the right member of (8.14), onto $P(T(z))$. The kernel of that map is isomorphic to the cokernel of column one in Fig. 8.4, which we denote by X''_z. From the left member of (8.14), there is an epic $k[z]$-linear map from the pole module of (5.3) onto X_M^T, with kernel isomorphic to

$$\frac{(\ker T(z)+T^{-1}Y[z] \cap M^{-1}U[z]) \cap R[z]}{T^{-1}Y[z] \cap M^{-1}U[z] \cap R[z]} = X'_k. \tag{8.15}$$

With the aid of the foregoing pieces of information, we can state the analogue of Theorem 10A for the case of Z'_Γ.

THEOREM 10B. *Let* $\ker M(z) \subset \ker T(z)$, *and let* Z'_Γ *be the finitely generated factor module of the matching zero module* Z_Γ. *Then there exist* $k[z]$-*modules* Z_M^T, Z'_k, X_M^T, X'_k, Z''_z, *and* X''_z, *and appropriate* $k[z]$-*linear maps, such that the following five short sequences are exact*:

$$0 \to X'_k \to P\left(\begin{bmatrix} T(z) \\ M(z) \end{bmatrix}\right) \to X_M^T \to 0, \tag{8.16a}$$

$$0 \to X''_z \to X_M^T \to P(T(z)) \to 0, \tag{8.16b}$$

$$0 \to Z'_k \to Z\left(\begin{bmatrix} T(z) \\ M(z) \end{bmatrix}\right) \to Z_M^T \to 0, \tag{8.16c}$$

$$0 \to Z_M^T \to Z(T(z)) \to Z''_z \to 0, \tag{8.16d}$$

$$0 \to X''_z \to Z'_\Gamma \to Z''_z \to 0. \tag{8.16e}$$

Proof. Equations (8.16a)-(8.16d) have already been established. Equation (8.16e) follows from the fact that the cokernels of the columns, including the induced, in Fig. 8.4 must fit into an exact sequence.

Discussion of Theorem 10B. The known quantities in (8.16) are the pole and zero modules, finitely generated and torsion, of (5.3) and of $T(z)$, and the pair X'_k and Z'_k, from (8.15) and (8.12), respectively. From these, we obtain X_M^T and Z_M^T, together with X''_z and Z''_z, and finally Z'_Γ. If $\ker M(z) = \ker T(z)$, then X'_k and Z'_k vanish, with resulting simplifications.

Although the results of §§ 2–8, Theorems 1–10, are stated for the ring $k[z]$, they hold in much more general situations. Section 9 describes some of these generalizations.

9. Extensions of the theory. For readability, we have chosen the ring $k[z]$ and the quotient field $k(z)$ in which to state the results of §§ 2–8. However, all these results remain true in more general contexts, sometimes with a modest re-interpretation of physical meaning. As an example, we can replace the subring $k[z]$ of $k(z)$ with any subring O which contains the base field k, which has quotient field $k(z)$, and which is a principal ideal domain. Such rings include localizations of $k[z]$ and discrete valuation rings. An important instance of the latter is O_∞, the subring of proper transfer functions. More illustrations of rings which are useful in system theory arise from forming the intersection of discrete valuation rings. One illustration here is the ring O_{ps} of transfer functions which are proper and stable, when $k = R$, the real numbers.

When $k[z]$ is replaced by O, (2.1) must be changed to

$$\Omega_O R = O \otimes_k R, \tag{9.1}$$

$$\Omega_O U = O \otimes_k U, \tag{9.2}$$

$$\Omega_O Y = O \otimes_k Y. \tag{9.3}$$

Elsewhere, simply replace $U[z]$ by $\Omega_O U$, and so forth.

In fact, the ideas of the paper also extend to systems defined over more general rings, with appropriate assumptions [18].

10. Conclusions. We have studied the constraints imposed upon the zeros of $k(z)$-linear maps $P(z): U(z) \to Y(z)$ and $M(z): R(z) \to U(z)$ by reason of the fact that they satisfy the model matching equation $T(z) = P(z)M(z)$, in which $T(z): R(z) \to Y(z)$ is $k(z)$-linear as well. This study was inspired by the work of Conte, Perdon, and Wyman [7] on fixed poles in the solution to this same equation. Yet it turns out that the situation for zeros differs markedly from that for poles. Indeed, Theorem 7 indicates that the usual type of zero—which we may think of intuitively as a state-space zero—does not have an intrinsic constraint imposed upon it from its appearance in a solution $P(z)$ or $M(z)$ to the equation. Rather, it is the extended zero, embodying the classical notion of zero together with the novel constructs of divisible or free zeros, which undergoes such constraint. Moreover, essential solutions—in which the constraint is achieved with equality—need not always exist; and this constitutes yet another distinction from the case for poles.

Because of the centrality of the model matching equation in many constructions of feedback systems and communications, we conclude that extended zeros may be just as important as, or perhaps even more important than, the classical or state-space type zeros. Accordingly, they appear to represent a fruitful topic for future research.

Finally these results are a consequence of the systematic use of module theory, which presents a convenient and natural framework for study of poles and zeros. Indeed, the existence and role of extended zeros have been discovered only as a result of its use.

REFERENCES

[1] J. L. Massey and M. K. Sain, *Inverses of linear sequential circuits*, in Key Papers in the Development of Coding Theory, E.R. Berlekamp, ed., IEEE Press, New York, 1974, pp. 205–212.

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

[3] S.-H. Wang and E. J. Davison, *A minimization algorithm for the design of linear multivariable systems*, IEEE Trans. Automat. Control, 18 (1973), pp. 220–225.

[4] A. S. MORSE, *Structure and design of linear model following systems*, IEEE Trans. Automat. Control, 18 (1973), pp. 346-353.

[5] W. A. WOLOVICH, P. J. ANTSAKLIS, AND H. ELLIOTT, *On the stability of solutions to minimal and nonminimal design problems*, IEEE Trans. Automat. Control, 22 (1977), pp. 88-94.

[6] J. HAMMER AND M. HEYMANN, *Causal factorizations and linear feedback*, SIAM J. Control Optim., 19 (1981), pp. 445-468.

[7] G. CONTE, A. M. PERDON, AND B. F. WYMAN, *Fixed poles in transfer function equations*, SIAM J. Control Optim., 26 (1988), pp. 356-368.

[8] B. F. WYMAN AND M. K. SAIN, *The zero module and essential inverse systems*, IEEE Trans. Circuits Systems, 27 (1981), pp. 112-126.

[9] ———, *On the design of pole modules for inverse systems*, IEEE Trans. Circuits Systems, 31 (1985), pp. 977-988.

[10] E. EMRE, *Generalized model matching and (F, G)-invariant submodules for linear systems over rings*, Linear Algebra Appl., 50 (1983), pp. 133-166.

[11] P. KHARGONEKAR AND E. EMRE, *Further results on polynomial characterizations of (F, G)-invariant and reachability subspaces*, IEEE Trans. Automat. Control, 27 (1982), pp. 352-366.

[12] E. EMRE AND M. L. J. HAUTUS, *A polynomial characterization of (A, B)-invariant and reachability subspaces*, SIAM J. Control Optim., 18 (1980), pp. 420-436.

[13] A. C. ANTOULAS, *On the dynamic cover problem*, Linear Algebra Appl., 50 (1983).

[14] B. F. WYMAN AND M. K. SAIN, *On the zeros of a minimal realization*, Linear Algebra Appl., 50 (1983), pp. 621-637.

[15] ———, *Exact sequences for pole-zero cancellation*, in Proc. Internat. Symposium on the Mathematical Theory of Networks and Systems, August 1981, pp. 278-280.

[16] G. CONTE AND A. M. PERDON, *Zeros of cascade compositions*, in Frequency Domain and State Space Methods for Linear Systems, C. I. Byrnes and A. Lindquist, eds., Elsevier, North-Holland, 1986, pp. 23-34.

[17] M. K. SAIN, *Introduction to Algebraic System Theory*, Chap. 7, Academic Press, New York, 1981.

[18] G. CONTE AND A. M. PERDON, *An algebraic notion of zeros for systems over rings*, in Proc. Internat. Symposium on the Mathematical Theory of Networks and Systems, 1983.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 3, pp. 594-608, May 1991

OPTIMAL DAMPING CONTROL AND NONLINEAR ELLIPTIC SYSTEMS*

SRDJAN STOJANOVIC†

Abstract. Optimal control for an elliptic equation when the control is the zero-order coefficient of the differential operator is considered and an optimality system is derived. Under certain assumptions, the problem is solved completely by giving the uniqueness and the constructive existence proof for the nonlinear optimality system. An extension to the case when the state equation is a particular elliptic system is considered as well.

Key words. optimal control, nonlinear optimality system, constructive existence proof

AMS(MOS) subject classifications. 49B22, 35J60

Introduction. We will consider an optimal control problem in which the state of the system is defined as the unique solution of an elliptic equation and where control is the zero-order coefficient of the differential operator (cf. [2], [9]). Also, an extension to a particular elliptic system is given.

After proving the existence of an optimal control, deriving the optimality system, and proving uniqueness for the optimality system (which happens to be a system of two nonlinear elliptic equations), we show how to solve that system; i.e., we give a *constructive* existence theorem.

When constructing a solution, the difficulty is that uniform estimates, i.e., compactness, i.e., convergence for a subsequence of some natural iteration, is not enough to solve the system (cf. [5]). We must invent a procedure with a property that the full sequence converges (see Theorem 4.1). Also, uniqueness for the optimality system is important not only because it provides uniqueness for the optimal control, and because it guarantees that the solution of the optimality system *is* an optimal control, but because the uniqueness for one closely related system (Lemma 4.1) guarantees the convergence of the approximating sequence (see § 4).

We present one computational example as well. It seems worth mentioning that numerical experiments suggest that somewhat restrictive assumptions of Theorem 4.1, which guarantee the convergence of the constructive scheme, could be relaxed (see Remark 5.1).

1. Optimal damping problem.

Let Ω be a bounded domain of $\mathbb{R}^n$ with $\mathbf{C}^{1,1}$ boundary. Let

$$f \in \mathbf{L}^q(\Omega), \quad q \geqq 2, \quad q > \frac{n}{2}, \tag{1.1}$$

and

$$\lambda \in \mathbb{R}_+ = \{\beta \mid \beta \in \mathbb{R}, \beta \geqq 0\} \tag{1.2}$$

be given. For any $s \geqq 1$, set

$$\mathbf{L}_+^s(\Omega) = \{c \mid c \in \mathbf{L}^s(\Omega), c \geqq 0 \text{ a.e. in } \Omega\}. \tag{1.3}$$

* Received by the editors August 10, 1988; accepted for publication (in revised form) April 27, 1990. This research was supported in part by Martin Marietta Energy Systems, Incorporated contract 19XSC136V, which is under contract DE-AC05-840R21400 with the United States Department of Energy.

† Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio 45221-0025.

For any $c \in \mathbf{L}^2_+(\Omega)$, we define $y = y(c)$ as a solution of (see [3])

$$(1.4)\qquad \begin{aligned} &-\Delta y + \lambda y + cy = f \quad \text{a.e. in } \Omega,\\ &y \in \mathbf{W}^{2,2}(\Omega) \cap \mathbf{H}^1_0(\Omega) \cap \mathbf{L}^\infty(\Omega). \end{aligned}$$

Remark 1.1. Results of this paper can be formulated for a general elliptic operator (with appropriate assumptions imposed) instead of $-\Delta$.

We observe that

(1.5) $\quad \|y(c)\|_{\mathbf{L}^\infty(\Omega)} \leqq \text{const.}$ where const. does not depend on $c \in \mathbf{L}^2_+(\Omega)$ and $\lambda \in \mathbb{R}_+$.

Next, for (q as in (1.1)

$$(1.6)\qquad y_d \in \mathbf{L}^q(\Omega)$$

and $N > 0$ given, we define the cost functional $\mathbf{J}$ by

$$(1.7)\qquad \mathbf{J}(c) = \frac{1}{2}\int_\Omega (y(c) - y_d)^2\, dx + \frac{N}{2}\int_\Omega (c)^2\, dx,$$

and we ask the question:

(1.8) How do we find an optimal control?

That is, how do we find (if it exists) a (damping) control $c_0 \in \mathbf{L}^2_+(\Omega)$ such that

$$(1.9)\qquad \mathbf{J}(c_0) = \inf_{c \in \mathbf{L}^2_+(\Omega)} \mathbf{J}(c).$$

Under certain assumptions, we will give a reasonably complete answer to the question (1.8). To start with, we observe that an optimal control always exists.

LEMMA 1.1. *There exists an optimal control.*

Proof. The proof is standard. Take a minimizing sequence $\{(c_n, y(c_n))\}$. Then $\|c_n\|_{\mathbf{L}^2(\Omega)}$ and $\|y(c_n)\|_{\mathbf{L}^\infty(\Omega)}$ are bounded sequences; hence, by elliptic estimates (see [3]), $\|y(c_n)\|_{\mathbf{W}^{2,2}(\Omega)}$ is bounded. Then there exist $c_0 \in \mathbf{L}^2_+(\Omega)$ and $y_0 \in \mathbf{L}^\infty(\Omega) \cap \mathbf{W}^{2,2}(\Omega)$, such that, for a subsequence,

$$\begin{aligned} c_n &\to c_0 \quad \text{weakly in } \mathbf{L}^2(\Omega),\\ y(c_n) &\to y_0 \quad \text{strongly in } \mathbf{H}^1_0(\Omega). \end{aligned}$$

Also, for any n we have

$$\int \nabla y(c_n) \cdot \nabla \varphi + \lambda y(c_n)\varphi + c_n y(c_n)\varphi = \int f\varphi \quad \forall \varphi \in \mathbf{H}^1_0(\Omega) \cap \mathbf{L}^\infty(\Omega).$$

Hence, passing the limit we conclude that

$$y_0 = y(c_0).$$

Also, by the lower semicontinuity of the cost functional, (c_0, y_0) is an optimal pair. □

2. Derivation of the optimality system. In this section we derive the necessary conditions, i.e., the optimality system, for an optimal pair (c, y). Two results are given in the next theorem. The first result holds under the general conditions already introduced. The second result holds under the condition that

$$(2.1)\qquad f \geqq 0 \quad \text{a.e. in } \Omega,$$

and

$$(2.2)\qquad y_d \leqq 0 \quad \text{a.e. in } \Omega.$$

THEOREM 3.1. (a) *Under the assumptions* (1.1), (1.2), *and* (1.6), *for any optimal pair* (c, y), *there exists* p *satisfying the system*

$$\begin{aligned}
&-\Delta y+\lambda y+cy=f \quad \text{a.e. in } \Omega, \quad y\in \mathbf{W}^{2,2}(\Omega)\cap \mathbf{H}_0^1(\Omega)\cap \mathbf{L}^\infty(\Omega),\\
&-\Delta p+\lambda p+cp-y=-y_d \quad \text{a.e. in } \Omega, \quad p\in \mathbf{W}^{2,2}(\Omega)\cap \mathbf{H}_0^1(\Omega)\cap \mathbf{L}^\infty(\Omega),\\
&c=\frac{1}{N}py \quad \text{a.e. in } \Omega\cap\{c>0\}, \qquad py\leqq 0 \quad \text{a.e. in } \Omega\cap\{c=0\}.
\end{aligned}\tag{2.3}$$

(b) *Under the assumptions* (1.1), (1.2), (1.6), (2.1), *and* (2.2), *for any optimal pair* (c, y), *there exists* p *satisfying the following nonlinear elliptic system*:

$$\begin{aligned}
&-\Delta y+\lambda y+\frac{1}{N}py^2=f \quad \text{a.e. in } \Omega, y=0, \quad \text{at } \partial\Omega, \quad y\in \mathbf{W}^{2,q}(\Omega),\\
&-\Delta p+\lambda p+\frac{1}{N}yp^2-y=-y_d \quad \text{a.e. in } \Omega, p=0 \quad \text{at } \partial\Omega, \quad p\in \mathbf{W}^{2,q}(\Omega),
\end{aligned}\tag{2.4}$$

and $c=py/N$. *Moreover*, $y\geqq 0$ *and* $p\geqq 0$.

Proof. It is not difficult to see (using standard arguments (cf. [1], [6])) that the mapping $\mathbf{L}^2_+(\Omega)\ni c\mapsto y(c)\in \mathbf{L}^2(\Omega)$ is differentiable in the following sense:

$$\frac{y(c+\beta\bar{c})-y(c)}{\beta}\to z \quad \text{strongly in } \mathbf{L}^2(\Omega), \quad \text{as } \beta\to 0,\tag{2.5}$$

for any $c\in \mathbf{L}^2_+(\Omega)$ and $\bar{c}\in \mathbf{L}^\infty(\Omega)$, such that $c+\beta\bar{c}\in \mathbf{L}^2_+(\Omega)$ (for $\beta\to 0$), and z is the solution to the equation

$$\begin{aligned}
&-\Delta z+\lambda z+cz=-y(c)\bar{c} \quad \text{a.e. in } \Omega,\\
&z\in \mathbf{W}^{2,2}(\Omega)\cap \mathbf{H}_0^1(\Omega)\cap \mathbf{L}^\infty(\Omega).
\end{aligned}\tag{2.6}$$

Let c be an optimal control. For every $\bar{c}\in \mathbf{L}^\infty_+(\Omega)$ and $\beta>0$, we have $c+\beta\bar{c}\in \mathbf{L}^2_+(\Omega)$, and

$$\begin{aligned}
&\frac{1}{2}\int_\Omega (y(c+\beta\bar{c})-y_d)^2\,dx+\frac{N}{2}\int_\Omega (c+\beta\bar{c})^2\,dx\\
&\geqq\frac{1}{2}\int_\Omega (y(c)-y_d)^2\,dx+\frac{N}{2}\int_\Omega (c)^2\,dx.
\end{aligned}$$

Hence, dividing by β, and sending $\beta\downarrow 0$, we get

$$\int_\Omega [(y(c)-y_d)z+Nc\bar{c}]\geqq 0.\tag{2.7}$$

Let p be a solution to the equation

$$\begin{aligned}
&-\Delta p+\lambda p+cp=y(c)-y_d \quad \text{a.e. in } \Omega,\\
&p\in \mathbf{W}^{2,2}(\Omega)\cap \mathbf{H}_0^1(\Omega)\cap \mathbf{L}^\infty(\Omega).
\end{aligned}\tag{2.8}$$

From (2.7) and (2.8) we get

$$\int_\Omega [(-\Delta p+\lambda p+cp)z+Nc\bar{c}]\geqq 0.$$

Integrating by parts and using (2.6), we deduce $(y=y(c))$

$$\int_\Omega (Nc-py)\bar{c}\geqq 0 \quad \forall \bar{c}\in \mathbf{L}^\infty_+(\Omega).\tag{2.9}$$

More precisely, for $c \in \mathbf{L}^2_+(\Omega)$, consider a variation $\bar{c} \in \mathbf{L}^\infty(\Omega)$. Define $\bar{c}_\varepsilon$ by

$$\bar{c}_\varepsilon = \begin{cases} \bar{c} & \text{if } c > \varepsilon \|\bar{c}\|_{\mathbf{L}^\infty(\Omega)}, \\ 0 & \text{elsewhere.} \end{cases} \tag{2.10}$$

Then for any $\beta \in \mathbb{R}$, such that $|\beta| < \varepsilon$,

$$c + \beta \bar{c}_\varepsilon \in \mathbf{L}^2_+(\Omega),$$

and proceeding as before, we conclude that, now

$$\int_\Omega (Nc - py)\bar{c}_\varepsilon = 0. \tag{2.11}$$

Passing $\varepsilon \downarrow 0$, we conclude that

$$\int_{\Omega \cap \{c>0\}} (Nc - py)\bar{c} = 0 \quad \forall \bar{c} \in \mathbf{L}^\infty(\Omega). \tag{2.12}$$

Putting together (2.8), (2.9), and (2.12), we prove (a).

Now, if (2.1) and (2.2) hold, then we have that

$$p \geqq 0 \quad \text{in } \Omega, \qquad y \geqq 0 \quad \text{in } \Omega,$$

and using (a) we conclude that

$$py = 0 \quad \text{a.e. in } \Omega \cap \{c = 0\},$$

hence,

$$c = \frac{1}{N} py \quad \text{in } \Omega,$$

and the theorem is proved. □

3. Uniqueness for the optimality system. In addition to (2.1) and (2.2) we assume that

(3.1) λ is sufficiently large.

Remark 3.1. From (1.5), (1.6), and (2.4) we see that there is a uniform constant C_0, depending only on the data (moreover, independent of λ and N), such that, for any optimal pair (c, y) and any corresponding p, we have

$$\max\{\|y\|_{\mathbf{L}^\infty(\Omega)}, \|p\|_{\mathbf{L}^\infty(\Omega)}\} \leqq C_0. \tag{3.2}$$

Then, we can write (3.1) explicitly, for example, a.

$$\lambda \geqq \frac{1}{2}\left|\frac{2}{N}(C_0)^2 - 1\right|. \tag{3.1}^*$$

PROPOSITION 3.1. *Under the assumptions* (1.1), (1.6), (2.1), (2.2), *and* (3.1) *there can be only one positive solution of* (2.4).

Proof. Weak formulation of (2.4) reads as

$$\begin{aligned} &\int_\Omega \left[\nabla y \cdot \nabla \psi + \lambda y \psi + \frac{1}{N} p y^2 \psi + \nabla p \cdot \nabla \varphi + \lambda p \varphi + \frac{1}{N} y p^2 \varphi - y \varphi\right] \\ &\quad = \int_\Omega [f\psi - y_d \varphi] \quad \forall (\psi, \varphi) \in [\mathbf{H}^1_0(\Omega)]^2 \cap [\mathbf{L}^\infty(\Omega)]^2; \end{aligned} \tag{3.3}$$

hence,

$$-\int_\Omega \left[|\nabla(p-\bar{p})|^2+|\nabla(y-\bar{y})|^2+\lambda(p-\bar{p})^2+\lambda(y-\bar{y})^2 + \frac{1}{N}(p-\bar{p})^2 y(p+\bar{p})+\frac{1}{N}(y-\bar{y})^2 p(y+\bar{y})\right] + \int_\Omega (\bar{p}-p)(y-\bar{y})\left[\frac{1}{N}(\bar{p}^2+\bar{y}^2)-1\right]=0.$$

It follows, since $y(p+\bar{p})\geqq 0$ and $p(y+\bar{y})\geqq 0$, that $p=\bar{p}$ and $y=\bar{y}$, provided

$$(3.1)^{**} \qquad \left\|\frac{1}{N}(\bar{p}^2+\bar{y}^2)-1\right\|_{\mathbf{L}^\infty(\Omega)} \leqq 2\lambda. \qquad \square$$

4. Solution of the optimality system. The purpose of this section is to give a constructive proof of the existence of a positive solution of the optimality system (2.4). Of course, since the proof is constructive, it can be easily used as an *algorithm for numerical computations* of the unique solutions of (2.4), which provides us with the optimal control. An example is given in the subsequent section.

We start with a simple problem:

(4.1) Find $u\in\mathbf{W}^{2,q}(\Omega)$, $u\geqq 0$, such that $-\Delta u+\lambda u+bu^2=f$ a.e. in Ω, $u=0$ at $\partial\Omega$.

We assume that

$$(4.2) \qquad b\in\mathbf{L}^\infty_+(\Omega).$$

Then we have the following proposition.

PROPOSITION 4.1. *Under the assumptions* (1.1), (1.2), (2.1), *and* (4.2), *there exists a unique solution of the problem* (4.1).

Proof. We shall prove the existence first. Let us start from u_0, the solution of

$$(4.3) \qquad \begin{aligned} &-\Delta u_0+\lambda u_0=f \quad \text{a.e. in } \Omega,\\ &u_0=0 \quad \text{at } \partial\Omega. \end{aligned}$$

Observe that (2.1) implies that $u_0\geqq 0$. Now, choose constant $M\geqq 0$, such that

(4.4) $-\|b\|_{\mathbf{L}^\infty(\Omega)}v^2+Mv$ is an increasing function of v, for $v\in[0, \|u_0\|_{\mathbf{L}^\infty(\Omega)}]$.

Next, for $k\geqq 1$, assuming we already have u_{k-1}, we define u_k as a solution of

$$(4.5) \qquad \begin{aligned} &-\Delta u_k+(\lambda+M)u_k=f-b(u_{k-1})^2+Mu_{k-1} \quad \text{a.e. in } \Omega,\\ &u_k=0 \quad \text{at } \partial\Omega. \end{aligned}$$

Then, for example, when $k=1$,

$$-\Delta(u_0-u_1)+M(u_0-u_1)=b(u_0)^2\geqq 0 \quad \text{a.e. in } \Omega,$$

and, hence, $u_0\geqq u_1$. Similarly, using (4.4), we conclude that $u_{k-1}\geqq u_k$ for any $k\geqq 1$. We conclude that full sequence converges, i.e.,

$$(4.6) \qquad u_k(x)\to u(x) \quad \forall x\in\Omega.$$

Also, by elliptic estimates, we have

$$(4.7) \qquad \|u_k\|_{\mathbf{W}^{2,q}(\Omega)}\leqq \text{const}.$$

Using (4.6) and (4.7) we can easily pass to the limit in (4.5) to conclude that u is a solution of (4.1).

Uniqueness will follow from the following lemma. It is worth mentioning that the lemma is a consequence of the monotonicity of the operator in (4.1) on the set $\{u \geqq 0\}$.

LEMMA 4.1. *Under the previous assumptions the following comparison result holds for the problem* (4.1):

$$b^* \geqq b, \qquad f^* \leqq f \Rightarrow u^* \leqq u. \tag{4.8}$$

Proof. We have

$$-\Delta(u^* - u) + \lambda(u^* - u) + b^*(u^*)^2 - b(u)^2 = f^* - f \leqq 0;$$

hence,

$$-\Delta(u^* - u) + \lambda(u^* - u) + b[(u^*)^2 - (u)^2] \leqq 0,$$

i.e.,

$$-\Delta(u^* - u) + \lambda(u^* - u) + b(u^* + u)(u^* - u) \leqq 0,$$

and, since $b(u^* + u) \geqq 0$, we conclude that $u^* \leqq u$. □

Next, we study the following (intermediate) problem:

(4.9) Find $(y, p) \in [\mathbf{W}^{2,q}(\Omega)]^2$, $y \geqq 0$, $p \geqq 0$, such that

$$-\Delta y + \lambda y + \frac{1}{N} p y^2 = f \quad \text{a.e. in } \Omega, \quad y = 0, \quad \text{at } \partial\Omega,$$

$$-\Delta p + \lambda p + \frac{1}{N} y p^2 = g \quad \text{a.e. in } \Omega, \quad p = 0 \quad \text{at } \partial\Omega.$$

We will assume that (for q as in (1.1))

$$g \in \mathbf{L}_+^q(\Omega). \tag{4.10}$$

We have the following proposition.

PROPOSITION 4.2. *Under the assumptions* (1.1), (1.2), (2.1), *and* (4.10) *there exists a solution* (y, p) *of* (4.9), *with the following comparison property*:

$$g^* \leqq g \Rightarrow p^* \leqq p, \qquad y^* \geqq y. \tag{4.11}$$

Remark 4.1. We can prove uniqueness for problem (4.9) using the same argument as in the proof of Proposition 3.1. Nevertheless, for this, instead of (3.1)**, we would need

$$\left\| \frac{1}{N}(p^2 + y^2) \right\|_{\mathbf{L}^\infty(\Omega)} \leqq 2\lambda. \tag{4.12}$$

We confine ourselves to prove only the existence of a solution satisfying property (4.11). This will be enough for our purposes.

Proof of Proposition 4.2. We shall construct in parallel, solutions corresponding to g and g^*. Define y_1 and y_1^* as a solution of

$$-\Delta y + \lambda y = f \quad \text{a.e. in } \Omega, \quad y = 0 \quad \text{at } \partial\Omega. \tag{4.13}$$

Next, using Proposition 4.1, define p_1 and p_1^* as a solution of

$$\begin{aligned} &-\Delta p + \lambda p + \frac{1}{N} y_1 p^2 = g \qquad \text{a.e. in } \Omega, \quad p = 0 \quad \text{at } \partial\Omega, \\ &-\Delta p + \lambda p + \frac{1}{N} y_1^* p^2 = g^* \quad \text{a.e. in } \Omega, \quad p = 0, \quad \text{at } \partial\Omega, \end{aligned} \tag{4.14}$$

respectively. By Lemma 4.1 we know that $p_1 \geqq p_1^*$. Furthermore, define y_2 and y_2^* as a solution of

$$(4.15)\qquad \begin{aligned} &-\Delta y+\lambda y+\frac{1}{N}p_1 y^2=f \quad \text{a.e. in } \Omega, \quad y=0 \quad \text{at } \partial\Omega, \\ &-\Delta y+\lambda y+\frac{1}{N}p_1^* y^2=f \quad \text{a.e. in } \Omega, \quad y=0 \quad \text{at } \partial\Omega, \end{aligned}$$

respectively. By Lemma 4.1 we conclude that $y_2 \leqq y_1$, $y_2^* \leqq y_1^*$ and that $y_2 \leqq y_2^*$. Proceeding similarly, we construct two sequences $\{(y_k, p_k)\}$ and $\{(y_k^*, p_k^*)\}$ such that, for $k \geqq 1$,

$$y_k \leqq y_{k-1}, \quad p_k \geqq p_{k-1}, \quad y_k^* \leqq y_{k-1}^*, \quad p_k^* \geqq p_{k-1}^*, \quad y_k \leqq y_k^*, \qquad p_k \geqq p_k^*.$$

Since those two sequences are bounded in $\mathbf{L}^\infty(\Omega)$, we can pass to the limit obtaining pointwise limits y, p, y^*, p^* for the full sequences and, moreover,

$$y \leqq y^* \quad \text{and} \quad p \geqq p^*.$$

Finally, since we have also uniform $\mathbf{W}^{2,q}(\Omega)$-estimates, we can conclude easily that the limiting functions satisfy (4.9). □

Remark 4.2. $y \leqq y_1$, where y is from the preceding proposition, and y_1 is defined in (4.13).

We are ready now to solve the original problem. We consider the following iterative scheme.

Let y_0 be defined as a solution of

$$(4.16)\qquad -\Delta y_0+\lambda y_0=f \quad \text{a.e. in } \Omega, \quad y_0=0 \quad \text{at } \partial\Omega.$$

Define (y_k, p_k), for $k \leqq 1$, as a solution, constructed in Proposition 4.2, of

$$(4.17)\qquad \begin{aligned} &-\Delta y_k+\lambda y_k+\frac{1}{N}p_k y_k^2=f \quad \text{a.e. in } \Omega, \quad y_k=0 \quad \text{at } \partial\Omega, \\ &-\Delta p_k+\lambda p_k+\frac{1}{N}y_k p_k^2=y_{k-1}-y_d \quad \text{a.e. in } \Omega, \quad y_k=0 \quad \text{at } \partial\Omega. \end{aligned}$$

Then we have Theorem 4.1.

THEOREM 4.1. *Under the assumptions* (1.1), (1.6), (2.1), (2.2), *and* (3.1) (*or* (3.1)*), *the following holds*:

$$(4.18)\qquad y_k \to y, \quad p_k \to p, \quad \text{as } k \to \infty,$$

weakly in $\mathbf{W}^{2,q}(\Omega)$, *for a full sequence. Hence,* (y, p) *is the unique positive solution of* (2.4), *and* $c_0 \equiv py/N$ *is the unique optimal control. Moreover,*

$$(4.19)\qquad y_{2k} \searrow y, \quad y_{2k+1} \nearrow y, \quad p_{2k} \nearrow p, \quad p_{2k+1} \searrow p \quad \text{as } k \to \infty.$$

Proof. From (4.16), (4.17) and Remark 4.2, we conclude that

$$y_0 \geqq y_2.$$

Hence, by the Proposition 4.2, we have

$$p_1 \geqq p_3, \qquad y_1 \leqq y_3,$$

which implies, in the same manner, that

$$p_2 \leqq p_4, \quad y_2 \geqq y_4, \qquad p_3 \geqq p_5, \quad y_3 \leqq y_5,$$

and so on. We conclude that

$$y_0 \geqq y_2 \geqq y_4 \geqq \cdots, \qquad y_1 \leqq y_3 \leqq y_5 \leqq \cdots,$$
$$p_2 \leqq p_4 \leqq p_6 \leqq \cdots, \qquad p_1 \geqq p_3 \geqq p_5 \geqq \cdots.$$

Hence, there are functions $\bar{y}$, $\underline{y}$, $\bar{p}$, $\underline{p}$, such that

$$y_{2k} \searrow \bar{y}, \quad y_{2k+1} \nearrow \underline{y}, \quad p_{2k} \nearrow \underline{p}, \quad p_{2k+1} \searrow \bar{p}, \tag{4.20}$$

and also, by the elliptic estimates, weakly in $\mathbf{W}^{2,q}(\Omega)$. Furthermore, we have

$$-\Delta y_{2k} + \lambda y_{2k} + \frac{1}{N} p_{2k}(y_{2k})^2 = f \quad \text{a.e. in } \Omega,$$

$$-\Delta p_{2k} + \lambda p_{2k} + \frac{1}{N} y_{2k}(p_{2k})^2 = y_{2k-1} - y_d \quad \text{a.e. in } \Omega,$$

and

$$-\Delta y_{2k+1} + \lambda y_{2k+1} + \frac{1}{N} p_{2k+1}(y_{2k+1})^2 = f \quad \text{a.e. in } \Omega,$$

$$-\Delta p_{2k+1} + \lambda p_{2k+1} + \frac{1}{N} y_{2k+1}(p_{2k+1})^2 = y_{2k} - y_d \quad \text{a.e. in } \Omega.$$

Passing to the limit, we conclude that $(\bar{y}, \underline{p}, \underline{y}, \bar{p})$ is a solution of

$$\begin{aligned}
&-\Delta \bar{y} + \lambda \bar{y} + \frac{1}{N} \underline{p} \bar{y}^2 = f \quad \text{a.e. in } \Omega, \quad \bar{y} = 0 \quad \text{at } \partial\Omega,\\
&-\Delta \underline{p} + \lambda \underline{p} + \frac{1}{N} \bar{y} \underline{p}^2 - \underline{y} = -y_d \quad \text{a.e. in } \Omega, \quad \underline{p} = 0 \quad \text{at } \partial\Omega,\\
&-\Delta \underline{y} + \lambda \underline{y} + \frac{1}{N} \bar{p} \underline{y}^2 = f \quad \text{a.e. in } \Omega, \quad \underline{y} = 0 \quad \text{at } \partial\Omega,\\
&-\Delta \bar{p} + \lambda \bar{p} + \frac{1}{N} \underline{y} \bar{p}^2 - \bar{y} = -y_d \quad \text{a.e. in } \Omega, \quad \bar{p} = 0 \quad \text{at } \partial\Omega.
\end{aligned} \tag{4.21}$$

Then, by the inspection, we see that $(\underline{y}, \bar{p}, \bar{y}, \underline{p})$ is a solution of (4.21) as well. So, if we show that, in an appropriate class (in a class where $(\bar{y}, \underline{p}, \underline{y}, \bar{p})$ and $(\underline{y}, \bar{p}, \bar{y}, \underline{p})$ belong), the uniqueness holds for problem (4.21), we can conclude that

$$\underline{y} = \bar{y}, \qquad \underline{p} = \bar{p},$$

and the theorem is proved. The following lemma addresses this last question.

LEMMA 4.2. *Suppose* $(\bar{y}, \underline{p}, \underline{y}, \bar{p})$ *and* $(\bar{l}, \underline{s}, \underline{l}, \bar{s})$ *are two positive solutions of* (4.21), *and suppose that*

$$\left\| \frac{1}{N}(\bar{l}^2 + \underline{s}^2) - 1 \right\|_{\mathbf{L}^\infty(\Omega)} \leqq 2\lambda, \tag{4.22}$$

and that

$$\left\| \frac{1}{N}(\underline{l}^2 + \bar{s}^2) - 1 \right\|_{\mathbf{L}^\infty(\Omega)} \leqq 2\lambda. \tag{4.23}$$

Then,

$$(\bar{y}, \underline{p}, \underline{y}, \bar{p}) = (\bar{l}, \underline{s}, \underline{l}, \bar{s}).$$

Proof. The proof is similar to the proof of Proposition 3.1. The variational formulation of (4.21) reads as

$$\int_\Omega \Big[\nabla\bar{y}\cdot\nabla\varphi_1+\lambda\bar{y}\varphi_1+\frac{1}{N}\underline{p}\bar{y}^2\varphi_1+\nabla\underline{p}\cdot\nabla\varphi_2+\lambda\underline{p}\varphi_2+\frac{1}{N}\bar{y}\underline{p}^2\varphi_2-\underline{y}\varphi_2$$

$$+\nabla\underline{y}\cdot\nabla\varphi_3+\lambda\underline{y}\varphi_3+\frac{1}{N}\bar{p}\underline{y}^2\varphi_3+\nabla\bar{p}\cdot\nabla\varphi_4+\lambda\bar{p}\varphi_4+\frac{1}{N}\underline{y}\bar{p}^2\varphi_4-\bar{y}\varphi_4\Big] \tag{4.24}$$

$$=\int_\Omega [f\varphi_1-y_d\varphi_2+f\varphi_3-y_d\varphi_4] \quad \forall(\varphi_1,\varphi_2,\varphi_3,\varphi_4)\in[\mathbf{H}_0^1(\Omega)]^4\cap[\mathbf{L}^\infty(\Omega)]^4.$$

Hence,

$$-\int_\Omega \Big[|\nabla(\bar{y}-\bar{l})|^2+|\nabla(\underline{p}-\underline{s})|^2+|\nabla(\underline{y}-\underline{l})|^2+|\nabla(\bar{p}-\bar{s})|^2$$

$$+\lambda((\bar{y}-\bar{l})^2+(\underline{p}-\underline{s})^2+(\underline{y}-\underline{l})^2+(\bar{p}-\bar{s})^2)$$

$$+\frac{1}{N}((\bar{l}-\bar{y})^2\underline{p}(\bar{y}+\bar{l})+(\underline{s}-\underline{p})^2\bar{y}(\underline{s}+\underline{p})+(\underline{l}-\underline{y})^2\bar{p}(\underline{l}+\underline{y})$$

$$+(\bar{s}-\bar{p})^2\underline{y}(\bar{s}+\bar{p}))\Big]$$

$$+\int_\Omega\Big[\Big(1-\frac{1}{N}(\bar{l}^2+\underline{s}^2)\Big)(\bar{y}-\bar{l})(\underline{p}-\underline{s})+\Big(1-\frac{1}{N}(\underline{l}^2+\bar{s}^2)\Big)(\underline{y}-\underline{l})(\bar{p}-\bar{s})\Big]=0.$$

The lemma follows since $\underline{p}(\bar{y}+\bar{l})$, $\bar{y}(\underline{s}+\underline{p})$, $\bar{p}(\underline{l}+\underline{y})$ and $\underline{y}(\bar{s}+\bar{p})$ are $\geqq 0$. Here we also use the assumptions (4.22) and (4.23). □

5. An example. In this section we present some computations done by implementing the algorithm developed in § 4.

Consider the state equation

$$\begin{aligned}&-\Delta u+cu=20(\sin(2\pi x)+1)\quad \text{in } \Omega=(0,1)\times(0.1),\\ &u=0\quad \text{on } \partial\Omega.\end{aligned} \tag{5.1}$$

The problem is to choose c, such that

$$\mathbf{J}(c)=\frac{1}{2}\int_\Omega [u(c)]^2\,dx+\frac{0.0005}{2}\int_\Omega c^2\,dx, \tag{5.2}$$

is minimized. Observe that

$$\lambda=0. \tag{5.3}$$

The optimality system is

$$\begin{aligned}&-\Delta u+\frac{1}{N}pu^2=20(\sin(2\pi x)+1)\quad \text{in } \Omega,\\ &-\Delta p+\frac{1}{N}up^2-u=0\quad \text{in } \Omega,\\ &u,p=0\quad \text{on } \partial\Omega.\end{aligned} \tag{5.4}$$

The program consists of several subroutines and a main program. Subroutine 1 is a solver for the problem (4.5). This is a single linear elliptic equation; hence, standard methods apply. We used piecewise linear finite elements. Subroutine 2 is a solver for the problem (4.1). It iterates Subroutine 1, as described in (4.5). Subroutine 3 is a solver for problem (4.9), using iteration defined in (4.13)-(4.15). Finally, the Main Program is a solver for the problem (2.4), using iteration (4.17).

Computations are done on Sun 386i workstation; the results are presented in Figs. 1–3. In Fig. 1, the solution of the state equation without any damping, i.e., $u(0)$, is given. In Fig. 2, the solution corresponding to the optimal damping $u(c)$ is presented. In Fig. 3, the optimal damping c is given. We observe that the convergence is fast: it takes only four to five iterations of the Subroutine 3 in the Main Program.

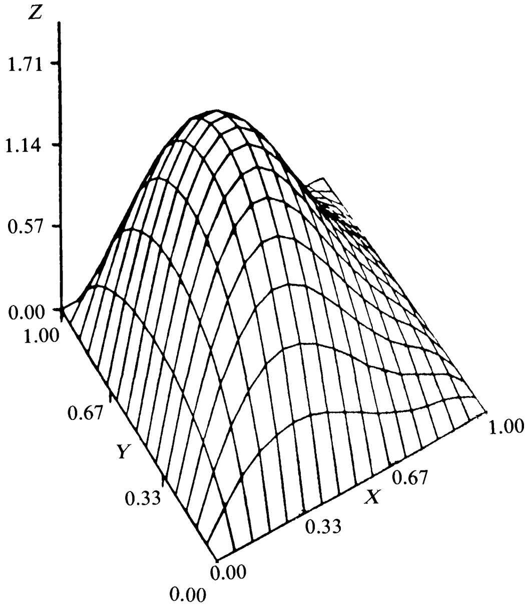


FIG. 1

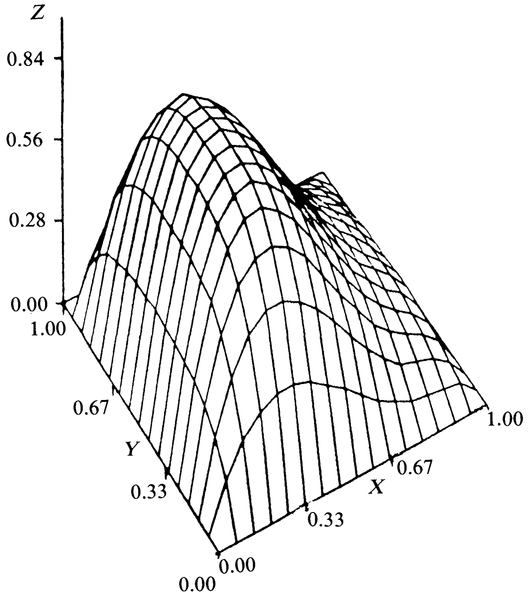


FIG. 2

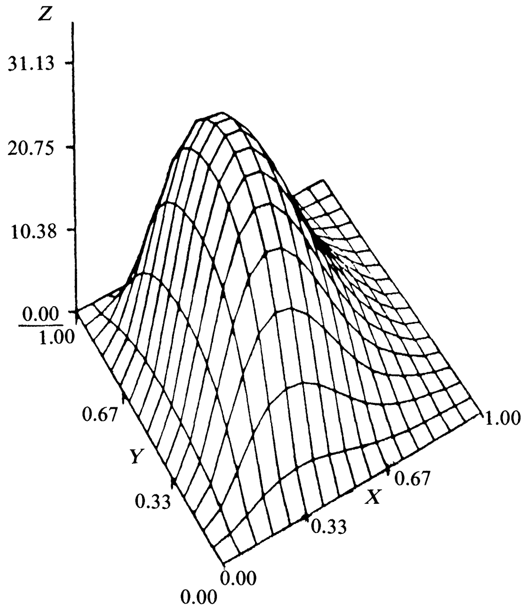


FIG. 3

Remark 5.1. In all numerical experiments that we have performed, the convergence has been recorded, even though λ was equal to zero. Hence, it seems reasonable to conjecture that Theorem 4.1 holds in that case as well. We observe that the largeness of λ was required to grant that

$$\underline{u} = \bar{u}, \qquad \underline{p} = \bar{p}.$$

6. An extension. Implicit damping control. Here we present, briefly, an extension to the case when the state equation is a particular elliptic system. Let

$$d \in \mathbf{L}^\infty_+(\Omega), \tag{6.1}$$

$$g \in \mathbf{L}^q(\Omega), \quad q \geqq 2, \quad q > \frac{n}{2}, \tag{6.2}$$

and

$$\lambda \in \mathbb{R}_+, \tag{6.3}$$

be given. For any $f \in \mathbf{L}^2_+(\Omega)$, we define $v = v(f)$ as a solution of the following equation:

$$\begin{gathered} -\Delta v + \lambda v + d[(-\Delta + \lambda)^{-1} f]v = g \quad \text{a.e. in } \Omega, \\ v \in \mathbf{W}^{2,2}(\Omega) \cap \mathbf{H}^1_0(\Omega) \cap \mathbf{L}^\infty(\Omega), \\ (-\Delta + \lambda)^{-1} f \in \mathbf{W}^{2,2}(\Omega) \cap \mathbf{H}^1_0(\Omega), \end{gathered} \tag{6.4}$$

or, more explicitly, we define $(u, v) = (u(f), v(f))$ as a solution of the following nonlinear system:

$$\begin{aligned} &-\Delta u + \lambda u = f \quad \text{a.e. in } \Omega, \\ &u \in \mathbf{W}^{2,2}(\Omega) \cap \mathbf{H}^1_0(\Omega), \\ &-\Delta v + \lambda v + duv = g \quad \text{a.e. in } \Omega, \\ &v \in \mathbf{W}^{2,2}(\Omega) \cap \mathbf{H}^1_0(\Omega) \cap \mathbf{L}^\infty(\Omega). \end{aligned} \tag{6.5}$$

Remark 6.1. Results of this section can be formulated for general elliptic operators, instead of $-\Delta$; more precisely, we can consider, instead of the state equation (1.5),

$$Av+\lambda v+d[(B+\lambda)^{-1}f]v=g,$$

where A and B are elliptic operators, with appropriate assumptions imposed.

Remark 6.2. When the first equation in (6.5) contains a zero-order term as in the second equation, it is natural to formulate, instead of the control problem considered here, a game problem. Then, the solution of the implicit damping problem can be viewed as a partial solution of the full game problem, which will be considered elsewhere (but under more restrictive assumptions).

It is easy to see that (6.5) has the unique solution. Also,

(6.6) $\|v(f)\|_{\mathbf{L}^{\infty}(\Omega)}\leqq$ const., const. does not depend on $\lambda\in\mathbb{R}_+$ and $f\in\mathbf{L}^2_+(\Omega)$.

Next, for (q as in (6.2))

$$v_d\in\mathbf{L}^q(\Omega) \tag{6.7}$$

and $N>0$ given, we define the cost functional $\mathbf{J}$ by

$$\mathbf{J}(f)=\frac{1}{2}\int_{\Omega}(v(f)-v_d)^2\,dx+\frac{N}{2}\int_{\Omega}f^2\,dx, \tag{6.8}$$

and, again, we look for an optimal control.

As before, it is not difficult to see that an optimal control always exists.

LEMMA 6.1. *There exists an optimal control.*

Again, we will need for the second part of the following theorem, the sign condition

$$g\geqq 0, \tag{6.9}$$

$$v_d\leqq 0. \tag{6.10}$$

THEOREM 6.1. (a) *Under the assumptions* (6.1), (6.2), (6.3), *and* (6.7), *for any optimal control f and associated solution to the state equation* (u, v), *there exists* (p_1, p_2) *satisfying the system*

$$\begin{aligned}
&-\Delta u+\lambda u=f \quad \textit{a.e. in } \Omega, \quad u\in\mathbf{W}^{2,2}(\Omega)\cap\mathbf{H}^1_0(\Omega),\\
&-\Delta v+\lambda v+duv=g \quad \textit{a.e. in } \Omega, \quad v\in\mathbf{W}^{2,2}(\Omega)\cap\mathbf{H}^1_0(\Omega)\cap\mathbf{L}^{\infty}(\Omega),\\
&-\Delta p_1+\lambda p_1-dvp_2=0 \quad \textit{a.e. in } \Omega, \quad p_1\in\mathbf{W}^{2,s}(\Omega), \quad p_1=0 \quad \textit{at } \partial\Omega,\\
&-\Delta p_2+\lambda p_2+dup_2-v=-v_d \quad \textit{a.e. in } \Omega, \quad p_2\in\mathbf{W}^{2,2}(\Omega)\cap\mathbf{H}^1_0(\Omega)\cap\mathbf{L}^{\infty}(\Omega),\\
&f=\frac{1}{N}p_1 \quad \textit{a.e. in } \Omega\cap\{f>0\}, \qquad p_1\leqq 0 \quad \textit{a.e. in } \Omega\cap\{f=0\},
\end{aligned} \tag{6.11}$$

for any $s<\infty$.

(b) *Under the assumptions* (6.1), (6.2), (6.3), (6.7), (6.9), *and* (6.10), *for any optimal control f and associated solution of the state equation* (u, v), *there exists* (p_1, p_2) *satisfying the following nonlinear elliptic system*:

$$\begin{aligned}
&-\Delta u+\lambda u-\frac{1}{N}p_1=0 \quad \textit{a.e. in } \Omega, \quad u\in\mathbf{W}^{2,s}(\Omega), \quad u=0 \quad \textit{at } \partial\Omega,\\
&-\Delta v+\lambda v+duv=g \quad \textit{a.e. in } \Omega, \quad v\in\mathbf{W}^{2,q}(\Omega), \quad v=0 \quad \textit{at } \partial\Omega,\\
&-\Delta p_1+\lambda p_1-dvp_2=0 \quad \textit{a.e. in } \Omega, \quad p_1\in\mathbf{W}^{2,s}(\Omega), \quad p_1=0 \quad \textit{at } \partial\Omega,\\
&-\Delta p_2+\lambda p_2+dup_2-v=-v_d \quad \textit{a.e. in } \Omega, \quad p_2\in\mathbf{W}^{2,q}(\Omega), \quad p_1=0 \quad \textit{at } \partial\Omega,
\end{aligned} \tag{6.12}$$

for any $s<\infty$, *and* $f=p_1/N$. *Moreover,* $u\geqq 0$, $v\geqq 0$, $p_1\geqq 0$, *and* $p_2\geqq 0$.

From (6.12), it is evident that there exists a constant C_0 depending only on data (moreover, independent of λ and N), such that

$$\|v\|_{L^\infty(\Omega)}, \|p_2\|_{L^\infty(\Omega)} \leqq C_0. \tag{6.13}$$

We assume that

$$\lambda \leqq \frac{1}{2} \max\left\{\frac{1}{N}, C_0\|d\|_{L^\infty(\Omega)}, 1\right\}. \tag{6.14}$$

LEMMA 6.2. *Under the assumptions* (6.1), (6.2), (6.3), (6.7), (6.9), (6.11), *and* (6.14), *there can be only one positive solution of* (6.12), *such that* (6.13) *holds.*

Finally, we construct solution of the optimality system. Consider the following auxiliary problem. Given $h \in L_+^\infty(\Omega)$, find $(u, v, p_1, p_2) \in W^{2,s}(\Omega) \times W^{2,q}(\Omega) \times W^{2,s}(\Omega) \times W^{2,q}(\Omega)$, for any $s < \infty$, such that

$$\begin{aligned} &-\Delta u + \lambda u = h \quad \text{a.e. in } \Omega, \quad u = 0 \quad \text{at } \partial\Omega,\\ &-\Delta v + \lambda v + duv = g \quad \text{a.e. in } \Omega, \quad v = 0 \quad \text{at } \partial\Omega,\\ &-\Delta p_1 + \lambda p_1 - dvp_2 = 0 \quad \text{a.e. in } \Omega, \quad p_1 = 0 \quad \text{at } \partial\Omega,\\ &-\Delta p_2 + \lambda p_2 + dup_2 - v = -v_d \quad \text{a.e. in } \Omega, \quad p_2 = 0 \quad \text{at } \partial\Omega. \end{aligned} \tag{6.15}$$

This is a trivial system. Indeed, $u \Rightarrow v \Rightarrow p_2 \Rightarrow p_1$. Also, we observe that

$$\bar{h} \geqq h \Rightarrow \bar{u} \geqq u \Rightarrow \bar{v} \leqq v \Rightarrow \bar{p}_2 \leqq p_2 \Rightarrow \bar{p}_1 \leqq p_1. \tag{6.16}$$

We are ready now to solve the original problem. Consider the following iterative scheme. Let $(u_0, v_0, p_{1,0}, p_{2,0})$ be a solution of

$$\begin{aligned} &u = 0 \quad \text{in } \Omega,\\ &-\Delta v + \lambda v = g \quad \text{a.e. in } \Omega, \quad v = 0 \quad \text{at } \partial\Omega,\\ &-\Delta p_1 + \lambda p_1 - dvp_2 = 0 \quad \text{a.e. in } \Omega, \quad p_1 = 0 \quad \text{at } \partial\Omega,\\ &-\Delta p_2 + \lambda p_2 + dup_2 - v = -v_d \quad \text{a.e. in } \Omega, \quad p_2 = 0 \quad \text{at } \partial\Omega. \end{aligned} \tag{6.17}$$

Define, by induction, for $k \geqq 1$, $(u_k, v_k, p_{1,k}, p_{2,k})$ as a solution of

$$\begin{aligned} &-\Delta u + \lambda u = \frac{1}{N} p_{1,k-1} \quad \text{a.e. in } \Omega, \quad u = 0 \quad \text{at } \partial\Omega,\\ &-\Delta v + \lambda v + duv = g \quad \text{a.e. in } \Omega, \quad v = 0 \quad \text{at } \partial\Omega,\\ &-\Delta p_1 + \lambda p_1 - dvp_2 = 0 \quad \text{a.e. in } \Omega, \quad p_1 = 0 \quad \text{at } \partial\Omega,\\ &-\Delta p_2 + \lambda p_2 + dup_2 - v = -v_d \quad \text{a.e. in } \Omega, \quad p_2 = 0 \quad \text{at } \partial\Omega. \end{aligned} \tag{6.18}$$

Then we have the following theorem.

THEOREM 6.2. *Under the assumptions* (6.1), (6.2), (6.3), (6.7), (6.9), (6.10), *and* (6.14), *the following holds*:

$$u_k \to u, \quad v_k \to v, \quad p_{1,k} \to p_1, \quad p_{2,k} \to p_2 \quad \text{as } k \to \infty, \tag{6.19}$$

weakly in $W^{2,q}(\Omega)$, *for a full sequence. Hence,* (u, v, p_1, p_2) *is the unique positive solution of* (6.12), *and* $f = p_1/N$ *is the unique optimal control. Moreover,*

$$\begin{aligned} &u_{2k} \nearrow u, \quad u_{2k+1} \searrow u, \quad v_{2k} \searrow v, \quad v_{2k+1} \nearrow v,\\ &p_{1,2k} \nearrow p_1, \quad p_{1,2k+1} \searrow p_1, \quad p_{2,2k} \searrow p_2, \quad p_{2,2k+1} \nearrow p_2, \end{aligned} \tag{6.20}$$

as $k \to \infty$.

Proof. It is not difficult to conclude from (6.16) that

$$u_{2k}\nearrow,\quad u_{2k+1}\searrow,\quad v_{2k}\searrow,\quad v_{2k+1}\nearrow,$$
$$p_{1,2k}\nearrow,\quad p_{1,2k+1}\searrow,\quad p_{2,2k}\searrow,\quad p_{2,2k+1}\nearrow. \tag{6.21}$$

Hence, there exists a $(\underline{u}, \bar{v}, \underline{p}_1, \bar{p}_2, \bar{u}, \underline{v}, \bar{p}_2, \underline{p}_2)$, such that

$$u_{2k}\nearrow \underline{u},\quad u_{2k+1}\searrow \bar{u},\quad v_{2k}\searrow \bar{v},\quad v_{2k+1}\nearrow \underline{v},$$
$$p_{1,2k}\nearrow \underline{p}_1,\quad p_{1,2k+1}\searrow \bar{p}_1,\quad p_{2,2k}\searrow \bar{p}_2,\quad p_{2,2k+1}\nearrow \underline{p}_2, \tag{6.22}$$

and also, by the elliptic estimates, weakly in $\mathbf{W}^{2,q}(\Omega)$ (actually, the u_i's and $p_{1,i}$'s converge weakly in $\mathbf{W}^{2,s}(\Omega)$, for any $s<\infty$). Furthermore, we have, for $k\geqq 1$,

$$-\Delta u_{2k}+\lambda u_{2k}-\frac{1}{N}p_{1,2k-1}=0 \quad \text{a.e. in } \Omega,\quad u_{2k}=0 \quad \text{at } \partial\Omega,$$
$$-\Delta v_{2k}+\lambda v_{2k}+du_{2k}v_{2k}=g \quad \text{a.e. in } \Omega,\ v_{2k}=0 \quad \text{at } \partial\Omega,$$
$$-\Delta p_{1,2k}+\lambda p_{1,2k}-dv_{2k}p_{2,2k}=0 \quad \text{a.e. in } \Omega,\quad p_{1,2k}=0 \quad \text{at } \partial\Omega,$$
$$-\Delta p_{2,2k}+\lambda p_{2,2k}+dup_{2,2k}-v_{2k}=-v_d \quad \text{a.e. in } \Omega,\quad p_{2,2k}=0 \quad \text{at } \partial\Omega,$$

and

$$-\Delta u_{2k+1}+\lambda u_{2k+1}-\frac{1}{N}p_{1,2k}=0 \quad \text{a.e. in } \Omega,\quad u_{2k+1}=0 \quad \text{at } \partial\Omega,$$
$$-\Delta v_{2k+1}+\lambda v_{2k+1}+du_{2k+1}v_{2k+1}=g \quad \text{a.e. in } \Omega,\quad v_{2k+1}=0 \quad \text{at } \partial\Omega,$$
$$-\Delta p_{1,2k+1}+\lambda p_{1,2k+1}-dv_{2k+1}p_{2,2k+1}=0 \quad \text{a.e. in } \Omega,\quad p_{1,2k+1}=0 \quad \text{at } \partial\Omega,$$
$$-\Delta p_{2,2k+1}+\lambda p_{2,2k+1}+dup_{2,2k+1}-v_{2k+1}=-v_d \quad \text{a.e. in } \Omega,\quad p_{2,2k+1}=0 \quad \text{at } \partial\Omega.$$

Passing to the limit, we conclude that $(\underline{u}, \bar{v}, \underline{p}_1, \bar{p}_2, \bar{u}, \underline{v}, \bar{p}_1, \underline{p}_2)$ is a solution of

$$-\Delta \underline{u}+\lambda \underline{u}-\frac{1}{N}\bar{p}_1=0 \quad \text{a.e. in } \Omega,\quad \underline{u}=0 \quad \text{at } \partial\Omega,$$
$$-\Delta \bar{v}+\lambda \bar{v}+d\underline{u}\bar{v}=g \quad \text{a.e. in } \Omega,\quad \bar{v}=0 \quad \text{at } \partial\Omega,$$
$$-\Delta \underline{p}_1+\lambda \underline{p}_1-d\bar{v}\bar{p}_2=0 \quad \text{a.e. in } \Omega,\quad \underline{p}_1=0 \quad \text{at } \partial\Omega,$$
$$-\Delta \bar{p}_2+\lambda \bar{p}_2+d\underline{u}\bar{p}_2-\bar{v}=-v_d \quad \text{a.e. in } \Omega,\quad \bar{p}_2=0 \quad \text{at } \partial\Omega.$$
$$-\Delta \bar{u}+\lambda \bar{u}-\frac{1}{N}\underline{p}_1=0 \quad \text{a.e. in } \Omega,\quad \bar{u}=0 \quad \text{at } \partial\Omega, \tag{6.23}$$
$$-\Delta \underline{v}+\lambda \underline{v}+d\bar{u}\underline{v}=g \quad \text{a.e. in } \Omega,\quad \underline{v}=0 \quad \text{at } \partial\Omega,$$
$$-\Delta \bar{p}_1+\lambda \bar{p}_1-d\bar{v}\underline{p}_2=0 \quad \text{a.e. in } \Omega,\quad \bar{p}_1=0 \quad \text{at } \partial\Omega,$$
$$-\Delta \underline{p}_2+\lambda \underline{p}_2+d\bar{u}\underline{p}_2-\underline{v}=-v_d \quad \text{a.e. in } \Omega,\quad \underline{p}_2=0 \quad \text{at } \partial\Omega.$$

Then, by the inspection, we see that $(\bar{u}, \underline{v}, \bar{p}_1, \underline{p}_2, \underline{u}, \bar{v}, \underline{p}_1, \bar{p}_2)$ is a solution, as well. But, quite similarly as in Lemma 4.2, we can show that, under the assumptions of the theorem, there can be only one solution of (6.23). Hence,

$$(\underline{u}, \bar{v}, \underline{p}_1, \bar{p}_2, \bar{u}, \underline{v}, \bar{p}_1, \underline{p}_2)=(\bar{u}, \underline{v}, \bar{p}_1, \underline{p}_2, \underline{u}, \bar{v}, \underline{p}_1, \bar{p}_2),$$

and we deduce (6.20). The theorem follows now easily. □

REFERENCES

[1] V. Barbu, *Optimal Control of Variational Inequalities*, Research Notes in Mathematics 100, Pitman, London, 1984.

[2] A. Friedman, *Nonlinear optimal control problems for parabolic equations*, SIAM J. Control Optim., 22 (1984), pp. 805-816.

[3] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[4] I. Lasiecka and R. Triggiani, eds., *Control Problems for Systems Described by Partial Differential Equations and Applications*, Lecture Notes in Control and Information Sciences, Vol. 97, Springer-Verlag, Berlin, 1987.

[5] A. Leung, *Systems of Nonlinear Partial Differential Equations and Applications*, Kluwer Academic Publishers, Dordrecht, the Netherlands, 1989.

[6] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, Berlin, 1971.

[7] ———, *Control of Distributed Singular Systems*, Gauthier-Villars, Paris, 1985.

[8] F. Mignot and J. P. Puel, *Optimal control in some variational inequalities*, SIAM J. Control Optim., 22 (1984), pp. 466-476.

[9] F. Murat, *Contre examples pour divers problemes ou le controle intervient dans les coefficients*, Ann. Mat. Pura Appl. (4), 112 (1976), pp. 49-68.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 3, pp. 609-622, May 1991

006

NONLINEAR NONINTERACTION WITH STABILITY BY DYNAMIC STATE FEEDBACK*

K. G. WAGNER†

Abstract. The dynamic feedback noninteraction with stability problem for square affine systems with a nonsingular decoupling matrix is investigated. First, a characterization of the class of dynamic feedbacks resulting in a noninteractive system is given. It is then shown that these feedbacks induce a certain subdynamic, which is determined by the given system alone and independent of the particular feedback used. The stability of this subdynamic is necessary for the achievement of noninteraction with stability. This condition extends an analogous result on the static feedback noninteraction with stability problem obtained recently by Isidori and Grizzle.

Key words. affine control systems, noninteraction with stability, dynamic noninteraction feedback, fixed dynamics

AMS(MOS) subject classifications. 93B10, 49E05

1. Introduction. Consider a square affine multi-input multi-output control system

$$(\Sigma)\qquad \begin{aligned} \dot{x} &= f(x) + \sum_{i=1}^{m} g_i(x)u_i, \\ y_i &= h_i(x), \qquad i = 1, \cdots, m, \quad m \geqq 2. \end{aligned}$$

Here the state x varies in $\mathbb{R}^n$, and the controls u_i and the outputs y_i are real-valued; f, g_i, h_i are assumed to be C^∞ mappings, defined on some open subset of $\mathbb{R}^n$. The following concepts are standard in control theory. (Σ) is called *noninteractive* if each output y_i is affected by the input u_i, but by no u_j, $j \neq i$. (Σ) is said to be *stable* at a reference point $\tilde{x}$ if $\tilde{x}$ is an asymptotically stable equilibrium point of the drift term $\dot{x} = f(x)$. The problem of the modification of (Σ) by feedback control such as to arrive at a noninteractive system is the *noninteracting control problem* [1], [2], [5], [7], [9]. If the feedback is also required to stabilize (Σ) locally at a given equilibrium of its drift term, we speak of the *noninteraction with stability problem* [3], [6], [9], [10]. The latter is the subject of the present paper. We discuss affine state feedbacks of the form

$$(\mathrm{F})\qquad \begin{aligned} \dot{\bar{x}} &= \bar{f}(x, \bar{x}) + \sum_{j=1}^{m} \bar{g}_j(x, \bar{x})v_j, \\ u_i &= \alpha_i(x, \bar{x}) + \sum_{j=1}^{m} \beta_{ij}(x, \bar{x})v_j, \qquad i = 1, \cdots, m, \end{aligned}$$

i.e., we include the possibility of a dynamic extension (state $\bar{x}$) and generally refer to (F) as a *dynamic feedback.* $v_1, \cdots, v_m$ are the new control variables. (F) contains the special case of a *static feedback* ($\dim \bar{x} = 0$).

In the discussion of the noninteraction with stability problem it is natural to consider only systems (Σ) that can be rendered separately noninteractive, respectively, stable by means of possibly different feedbacks. The question then is which further

* Received by the editors February 21, 1989; accepted for publication (in revised form) May 29, 1990. This research was supported by Deutsche Forschungsgemeinschaft Wa 583/1-1 and by the Università degli Studi di Roma "La Sapienza," Rome, Italy.

† Mathematisches Institut der Universität Würzburg, 8700 Würzburg, Germany. Most of this research was done while visiting the Dipartimento di Informatica e Sistemistica, Università "La Sapienza," Rome, Italy.

conditions are needed in order to achieve both objectives simultaneously. For the case where only static feedback is allowed, the crucial condition has been found recently by Isidori and Grizzle [6]. It generalizes an analogous result for linear systems due to Gilbert [3] and amounts to the requirement that a certain invariant subsystem defined by (Σ), the so-called P^* *dynamics*, must be a priori stable (see § 3).

The knowledge about the dynamic feedback noninteraction with stability problem is less complete. For linear systems, Wonham and Morse showed the very satisfactory result that no additional conditions at all need to be imposed [9], [10]. In nonlinear systems theory, however, the problem remained open until, in [6], an example was presented that demonstrated that it may well be impossible to overcome the obstruction of an unstable P^* dynamics even by dynamic feedback, in contrast to the linear case.

In this paper we consider systems for which the static feedback noninteracting control problem is solvable. We show that a general genuinely nonlinear phenomenon underlies the cited example. It turns out that the P^* dynamics contains another, generally lower-dimensional subdynamic, called the Δ_{mix} *dynamics*, which is invariant even under dynamic feedbacks. If this dynamics is unstable, noninteraction with stability cannot be achieved. For linear systems, the Δ_{mix} dynamics reduces to dimension zero (i.e., does not appear), in accordance with the Wonham-Morse theory. An example shows that also in the nonlinear case a system with an unstable P^* dynamics may still be rendered both noninteractive and stable via a suitable dynamic feedback, provided the smaller Δ_{mix} dynamics is stable.

Throughout this paper we extensively use some standard methods of geometric control theory that by now are quite common in the field. All the necessary background material can be found in [5]. Also, our notation follows that of [5] and [6].

2. Regular noninteraction feedbacks. The following criterion will be used extensively: if (Σ) is noninteractive, then (see [5])

(H1) For $i=1,\cdots,m$ and for any product D of the differential operators L_f, L_{g_i} we have

$$L_{g_j}Dh_i \equiv 0 \quad \text{for all } j \neq i \tag{2.1}$$

(here L denotes the Lie derivative). Throughout this paper we consider the system (Σ) locally near the reference point $x=0$. We assume

(H2) Each output h_i has a well-defined (local) *characteristic number* ρ_i at $x=0$ with respect to (Σ), i.e., for $0 \leqq k < \rho_i$ and each j: $L_{g_j}L_f^k h_i \equiv 0$, and for some j: $L_{g_j}L_f^{\rho_i}h_i(0) \neq 0$,

$$\text{i.e., when (H1) holds: } L_{g_i}L_f^{\rho_i}h_i(0) \neq 0. \tag{2.2}$$

Given (H2), (H1) is also sufficient for noninteraction at least if f, g_i, h_i are analytic [4], [5]. So the following definition is natural.

DEFINITION 2.1. (i) A feedback (F) transforming (Σ) into the system

$$(\Sigma^e) \qquad \begin{aligned} \dot{x}^e &= F(x^e) + \sum_{j=1}^{m} G_j(x^e)v_j, \\ y_i &= H_i(x^e) = h_i(x), \qquad i=1,\cdots,m, \end{aligned}$$

where $x^e = \binom{x}{\bar{x}}$ and

$$F(x^e) = \begin{pmatrix} f(x) \\ \bar{f}(x^e) \end{pmatrix} + \sum_{i=1}^{m} \begin{pmatrix} g_i(x) \\ 0 \end{pmatrix} \alpha_i(x^e), \tag{2.3a}$$

$$G_j(x^e) = \begin{pmatrix} 0 \\ \bar{g}_j(x^e) \end{pmatrix} + \sum_{i=1}^{m} \begin{pmatrix} g_i(x) \\ 0 \end{pmatrix} \beta_{ij}(x^e), \qquad 1 \leqq j \leqq m, \tag{2.3b}$$

is called a *noninteraction feedback* (for (Σ)) if for any i and any product D^e of factors L_F, L_{G_i}

$$L_{G_j} D^e h_i \equiv 0 \quad \text{for all } j \neq i. \tag{2.4}$$

(ii) If, in addition, each output h_i has a characteristic number σ_i at the origin $x^e = 0$ with respect to (Σ^e), then (F) is called a *regular noninteraction feedback.*

(iii) When dealing with systems (Σ) with $f(0) = 0$ we say that the feedback (F) *preserves the equilibrium* if $\bar{f}(0, 0) = 0$ and $\alpha_i(0, 0) = 0$ for $i = 1, \cdots, m$ (so $F(0) = 0$).

The object of this preparatory section is a characterization of the class of regular noninteraction feedbacks for those systems (Σ) that are already noninteractive.

So assume that (H1), (H2) hold for (Σ) and consider a feedback (F) and the resulting system (Σ^e). For notational convenience from now on we will often omit the argument x^e when a mapping $H(x^e)$ is referred to. If H happens to depend only on the x part of $x^e = (x, \bar{x})$, we emphasize this by writing $H(x)$. The following formulas are easily deducted from the definitions:

$$\begin{aligned} &L_F^k h_i = L_f^k h_i(x), \qquad 0 \leqq k \leqq \rho_i, \\ &L_F^{\rho_i+1} h_i = L_f^{\rho_i+1} h_i(x) + L_{g_i} L_f^{\rho_i} h_i(x) \cdot \alpha_i, \\ &L_{G_j} L_F^k h_i = L_{g_i} L_f^k h_i(x) \cdot \beta_{ij}, \qquad 0 \leqq k \leqq \rho_i \qquad (=0 \text{ for } k < \rho_i). \end{aligned} \tag{2.5}$$

If (F) is a noninteraction feedback, then (2.5), (2.4), (2.2) imply $\beta_{ij} \equiv 0$ for $i \neq j$; thus, (2.3b) reduces to

$$G_j = \begin{pmatrix} 0 \\ \bar{g}_j \end{pmatrix} + \begin{pmatrix} g_j \\ 0 \end{pmatrix} (x) \cdot \beta_{jj}, \qquad j = 1, \cdots, m. \tag{2.6}$$

In addition, if $\beta_{ii}(x^e)$ is not identically zero on a neighborhood of $x^e = 0$, (2.5) implies that the characteristic number σ_i of h_i at $x^e = 0$ with respect to (Σ^e) exists if and only if $\beta_{ii}(0) \neq 0$ and in this case, $\sigma_i = \rho_i$. This is case (i) of the following claim (2).

PROPOSITION 2.2. *Assume that* (Σ) *satisfies* (H1) *and* (H2). *Claim*:

(1) *A feedback* (F) *is a noninteraction feedback if and only if its coefficients* α_i, β_{ij} *exhibit the following* decoupling property:

$$\beta_{ij} \equiv 0 \quad \text{if } j \neq i, \tag{2.7a}$$

$$L_{G_j} D^e \alpha_i \equiv 0, \qquad L_{G_j} D^e \beta_{ii} \equiv 0 \tag{2.7b}$$

for all $i = 1, \cdots, m$, *all* $j \neq i$ *and any product* D^e *of factors* L_F, L_{G_i}.

(2) (F) *is a regular noninteraction feedback if, in addition, for each* $i \in \{1, \cdots, m\}$ *one of the following statements holds*:

(i) $\beta_{ii}(0) \neq 0$, *or*

(ii) $\beta_{ii} \equiv 0$, *and* $\alpha_i(x^e)$ *has a characteristic number, say* $\tau_i - 1$ *at* $x^e = 0$ *with respect to* (Σ^e).

The characteristic number σ_i *of* h_i *at* $x^e = 0$ *with respect to* (Σ^e) *equals* ρ_i *in case* (i), *respectively,* $\rho_i + \tau_i$ *in case* (ii).

Sketch of proof. We may assume (2.7a). Let D^e be a product of factors L_F, L_{G_i}. The idea of the proof is to establish the following relations between the relevant Lie derivatives of h_i, β_{ii}, α_i (to prove by a straightforward induction on the number of factors of D^e, starting with (2.5)):

$$D^e L_{G_i} L_F^{\rho_i} h_i = L_{g_i} L_f^{\rho_i} h_i(x) \cdot D^e \beta_{ii} + \sum_{\text{finite}} \phi_{k,D^e}(x) \cdot B_{k,D^e}, \tag{2.8}$$

$$D^e L_F^{\rho_i+1} h_i = L_{g_i} L_f^{\rho_i} h_i(x) \cdot D^e \alpha_i + \sum_{\text{finite}} \psi_{k,D^e}(x) \cdot A_{k,D^e}. \tag{2.9}$$

Here k is a summation index. Each function ϕ_{k,D^e}, ψ_{k,D^e} is of the form $DL_{g_i}L_f^{\rho_i}h_i$ or $DL_f^{\rho_i+1}h_i$ for some product D of factors L_f, L_{g_i}. Each B_{k,D^e}, A_{k,D^e} is a product of functions of the form $\bar{D}^e\beta_{ii}$ or $\bar{D}^e\alpha_i$, the $\bar{D}^e$ being products of factors L_F, L_{G_i}. The total number of operators L_F, L_{G_i} appearing in B_{k,D^e}, respectively, A_{k,D^e} is strictly smaller than the number of factors of D^e (harmless exception: for $D^e =$ identity an $A_{k,D^e} \equiv 1$ appears, cf. (2.5)).

Claim (1) can now be derived from (2.8), (2.9) by a similar induction argument. The $\beta_{ii} \equiv 0$ case of claim (2) results from the following auxiliary claim: If $\beta_{ii} \equiv 0$, $\tau \geqq 1$ and $L_{G_i}L_F^t\alpha_i \equiv 0$, for $0 \leqq t < \tau - 1$, then $L_{G_i}L_F^{\rho_i+\tau}h_i = L_{g_i}L_f^{\rho_i}h_i(x) \cdot L_{G_i}L_F^{\tau-1}\alpha_i$. (To derive for $\tau = 1$ from (2.5) and, for $\tau > 1$, from (2.9): let $D^e = L_F^{\tau-1}$, then apply L_{G_i} and note (2.6).) □

Remark. The decoupling condition (2.7b) can also be expressed in terms of first-order differential operators involving Lie brackets. An elementary calculation using the rule $L_{[C,D]} = L_CL_D - L_DL_C$ shows the following well-known equivalence for a function $\alpha(x^e)$:

$$L_{G_j}D^e\alpha \equiv 0$$

$$\text{for all } j \neq i \text{ and all products } D^e \text{ of factors } L_F,\ L_{G_i}$$

if and only if

$$L_B\alpha \equiv 0$$

$$\text{for all Lie products } B \text{ of } F, G_1, \cdots, G_m \tag{2.10}$$

$$\text{containing a factor } G_j, j \neq i.$$

The form (2.10) for the decoupling condition will be useful later.

Proposition 2.2 may be compared with a similar characterization of the class of static regular noninteraction feedbacks due to Ha and Gilbert [4]. Its essential part is the decoupling property for α_i, β_{ii} with respect to the original system (Σ) instead of (Σ^e). In contrast, our condition refers to (Σ^e) and is thus only an implicit one. Nevertheless, it will serve as an essential tool for the proofs of our main results.

3. Review: the P^* decomposition. Let (H2) hold and let $A(x)$ be the so-called *decoupling matrix* of (Σ), defined by its entries

$$a_{ij}(x) = L_{g_j}L_f^{\rho_i}h_i(x), \qquad 1 \leqq i, j \leqq m.$$

Throughout the rest of this paper we assume that a static regular noninteraction feedback for (Σ) exists locally near $x = 0$. This is the case if and only if (see [5])

(H3) $A(x)$ is nonsingular near $x = 0$.

The starting point for the ensuing discussion is the main result of [6] on the static feedback noninteraction with stability problem which we briefly review here. Let R^* be the smallest distribution containing $g_1, \cdots, g_m$ and invariant under $f, g_1, \cdots, g_m$. Following [5] we denote this by

$$R^* = \langle f, g_1, \cdots, g_m | \operatorname{sp}\{g_1, \cdots, g_m\}\rangle.$$

R^* is the so-called strong accessibility distribution of (Σ) [8], [5, p. 38]. (H2) implies that, for $i = 1, \cdots, m$, the maximal controlled-invariant distribution Δ_i^* contained in $\ker dh_i$ is well-defined, nonsingular, and involutive [5, p. 145]. We need the following additional regularity assumption:

(H4) (i) Near $x = 0$, R^* is nonsingular and finitely computable (cf. [5, § I.6]).

(ii) Near $x = 0$, the largest controllability distribution P_i^* contained in Δ_i^* is defined, nonsingular, and finitely computable for $i = 1, \cdots, m$ [5, § IV.5].

Hypothesis (H4) is known to hold for an open and dense set of reference points. What we actually need is that $x=0$ belongs to this set.

THEOREM 3.1 [6]. *Let* (H2), (H3), (H4) *hold for* (Σ). *Then there exists a coordinate system* $x=(x_1,\cdots,x_{m+2})$ *locally around* $x=0$, *each* x_i *possibly vector-valued, with the following properties*:

(1) $P_i^* = \mathrm{sp}\,\{\partial/\partial x_j : 1 \leqq j \leqq m+1, j \neq i\}$,

$$P^* := \bigcap_{i=1}^{m} P_i^* = \mathrm{sp}\left\{\frac{\partial}{\partial x_{m+1}}\right\}. \tag{3.1}$$

(2) *In the coordinates* x, *each noninteractive system* ($\tilde{\Sigma}$) *obtained from* (Σ) *by means of a static regular noninteraction feedback displays the following* P^* decomposed form:

$$\begin{aligned} \dot{x}_1 &= \tilde{f}_1(x_1, x_{m+2}) + \tilde{g}_{11}(x_1, x_{m+2})v_1, && y_1 = h_1(x_1, x_{m+2}) \\ &\vdots \qquad\qquad \vdots && \vdots \\ \dot{x}_m &= \tilde{f}_m(x_m, x_{m+2}) + \tilde{g}_{mm}(x_m, x_{m+2})v_m, && y_m = h_m(x_m, x_{m+2}) \\ \dot{x}_{m+1} &= \tilde{f}_{m+1}(x) + \sum_{i=1}^{m} \tilde{g}_{i,m+1}(x)v_i \\ \dot{x}_{m+2} &= \tilde{f}_{m+2}(x_{m+2}). \end{aligned} \tag{3.2}$$

In addition we have in terms of ($\tilde{\Sigma}$):

$$P_i^* = \langle \tilde{f}, \tilde{g}_1, \cdots, \tilde{g}_m \,|\, \mathrm{sp}\,\{\tilde{g}_j : 1 \leqq j \leqq m, j \neq i\}\rangle. \tag{3.3}$$

(3) *Now suppose* $f(0)=0$ *and consider in* (2) *only feedbacks that preserve the equilibrium, so* $\tilde{f}(0)=0$ *in* (3.2). *Then the map* $\tilde{f}_{m+1}(0,\cdots,0,x_{m+1},0)$ *is determined by the given system* (Σ) *alone. The drift term of* ($\tilde{\Sigma}$) *can be asymptotically stable at* $x=0$ *only if the* P^* fixed dynamics

$$\dot{x}_{m+1} = \tilde{f}_{m+1}(0,\cdots,0,x_{m+1},0) \tag{3.4}$$

associated with (Σ) *is asymptotically stable at* $x_{m+1}=0$.

Remark. If dynamic feedback is considered, the stability of the P^* fixed dynamics is no longer necessary for noninteraction with stability. For linear systems, this is clear in view of the Wonham-Morse theory mentioned in the Introduction. A nonlinear example will be provided later in § 6.

4. The mixed brackets distribution and the corresponding subsystem. Invariance properties. The result just described indicates that the P^* dynamics constitutes the essential obstruction in the static feedback noninteraction with stability problem. It is natural to ask how far we can reduce this P^* dynamics if dynamic feedback is allowed. To motivate the following discussion a bit, suppose for the moment that (Σ) is already noninteractive and consider the case (Σ) = ($\tilde{\Sigma}$) in Theorem 3.1. By (3.3), (3.1) it is clear that all those Lie products of $f, g_1, \cdots, g_m$ that contain two factors g_i, g_j with $i \neq j$ belong "canonically" to P^*. We refer to these products in the sequel as *mixed brackets.* Of course, P^* may be bigger. For instance, if $[f, g_1]$ and $[f, g_2]$ happen to be nonzero and linearly dependent, then these nonmixed brackets also belong to P^*. However, the existence of such additional elements of P^* depends on the validity of certain nontrivial relations in the Lie algebra Lie $\{f, g_1, \cdots, g_m\}$ that might be destroyed by means of a suitable dynamic extension. Thus, it is reasonable to conjecture that P^* and, consequently, the P^* subsystem have a "core" determined by the mixed brackets and this intuition is indeed correct. So let

Δ_{mix}, the *mixed brackets distribution* be the distribution generated by all the mixed brackets. While Δ_i^*, P_i^*, P^* are invariant, at least under invertible static feedbacks, there is no obvious reason for expecting any kind of feedback invariance of the mixed brackets distribution. Yet our main result states that Δ_{mix} does exhibit invariance, in a sense, with respect to noninteraction feedbacks.

THEOREM 4.1 (invariant projection property of Δ_{mix}). *Suppose that* (Σ) *satisfies hypotheses* (H1) *and* (H2). *Let* (Σ^e) *be any system obtained from* (Σ) *by means of a* (*dynamic*) *regular noninteraction feedback. Let* Δ_{mix}^e *be the mixed brackets distribution of* (Σ^e) *and* π_x, *the canonical projection from the tangent bundle of the* x^e *space onto that of the* x *space. Then*

$$\pi_x(\Delta_{\text{mix}}^e) = \Delta_{\text{mix}}.$$

The proof of this theorem is given in the next section.

COROLLARY 4.2. *If in Theorem* 4.1 *the feedback applied to* (Σ) *is static, then* $\Delta_{\text{mix}}^e = \Delta_{\text{mix}}$.

Remark 4.3. As a consequence, if we start with a system (Σ') that satisfies (H2) and (H3) but is not itself noninteractive, then any two systems (Σ_1), (Σ_2) obtained from (Σ') by means of a static regular noninteraction feedback will result in the same mixed brackets distribution Δ_{mix} (for (Σ_1), (Σ_2) can be transformed into each other via static regular noninteraction feedbacks). That is, this Δ_{mix} is in fact determined by the given system (Σ') alone and does not depend on the particular choice of its noninteractive counterpart (Σ_i).

We now use the mixed brackets distribution to derive a refinement of the P^* decomposed form (3.2). Suppose that $f(0) = 0$ and that (Σ) also satisfies (H4) in addition to (H1) and (H2), so Theorem 3.1 can be applied for $(\tilde{\Sigma}) = (\Sigma)$ (accordingly, we omit $\tilde{}$ and write u_i instead of v_i when we refer to the P^* decomposed form (3.2) of (Σ) itself). Suppose further that

(H5) Δ_{mix} is of constant dimension locally around $x = 0$.

Given (H5) it follows by the very definition of Δ_{mix} that Δ_{mix} is an involutive subdistribution of P^* and is invariant under $f, g_1, \cdots, g_m$. In particular, the differentials of the components of the coordinates $x_1, \cdots, x_m, x_{m+2}$ used in Theorem 3.1 annihilate Δ_{mix}. Now recall that during the construction of the coordinate system x (see [6, Lemma 4.1]) the x_{m+1} part is required only to complete the previously constructed $x_1, \cdots, x_m, x_{m+2}$ to a full coordinate system. So we can, without affecting the statements of Theorem 3.1, choose new coordinates x_{m+1} as follows. First, choose x_{m+1}^{**} such that the components of $dx_1, \cdots, dx_m, dx_{m+1}^{**}, dx_{m+2}$ constitute a local base for the annihilator of Δ_{mix} (this is possible by the Frobenius theorem since Δ_{mix} is involutive and nonsingular). Then adjoin further coordinates x_{m+1}^* so as to obtain a full coordinate system. Then

$$\Delta_{\text{mix}} = \text{sp}\left\{\frac{\partial}{\partial x_{m+1}^*}\right\}, \tag{4.1}$$

and since Δ_{mix} is invariant under $f, g_1, \cdots, g_m$, the x_{m+1}^* coordinates appear in (3.2) nowhere but in the $\dot{x}_{m+1}^*$ equation (cf. [5, Lemma I.4.3]). Consequently, the P^* dynamics (3.4) splits:

$$\dot{x}_{m+1}^* = f_{m+1}^*(0, \cdots, 0, x_{m+1}^*, x_{m+1}^{**}, 0),$$

$$\dot{x}_{m+1}^{**} = f_{m+1}^{**}(0, \cdots, 0, x_{m+1}^{**}, 0).$$

For this dynamics, $x^{**}_{m+1}=0$ defines an invariant manifold and the restriction on this manifold is given by

$$\dot{x}^*_{m+1}=f^*_{m+1}(0,\cdots,0,x^*_{m+1},0,0). \tag{4.2}$$

This is (in coordinates) the restriction of the drift term of (Σ) to the integral manifold of Δ_{mix} passing through $x=0$; therefore, we call it the Δ_{mix} *dynamics.* Since P^*, Δ_{mix}, and the P^* dynamics are the same after any static regular noninteraction feedback, so is the Δ_{mix} dynamics. Accordingly, we can, in analogy to Remark 4.3, associate a well-defined Δ_{mix} dynamics also to systems (Σ') that satisfy only (H3) instead of (H1), simply by executing first an arbitrary static regular noninteraction feedback and then using the resulting system (Σ).

The most remarkable feature of the Δ_{mix} dynamics is that, unlike the P^* dynamics, it also survives dynamic noninteraction feedbacks. The following theorem constitutes an analogue to Theorem 3.1(3).

THEOREM 4.4 (invariant projection property of the Δ_{mix} dynamics). *Suppose that* $f(0)=0$ *and that* (Σ) *satisfies* (H1), (H2), (H4), (H5). *Let* (Σ^e) *be a system obtained from* (Σ) *by means of a (dynamic) regular noninteraction feedback which preserves the equilibrium. Assume that the mixed brackets distribution* Δ^e_{mix} *of* (Σ^e) *is also constant dimensional near* $x^e=0$. *Claim*:

(1) *The drift term of* (Σ^e) *admits the restriction on the integral manifold of* Δ^e_{mix} *passing through* $x^e=0$.

(2) *This restriction, called the* Δ^e_{mix} *dynamics of* (Σ^e), *has a well-defined projection into the x space, and this projection coincides with the* Δ_{mix} *dynamics of* (Σ). *Consequently,* (Σ^e) *can be asymptotically stable at* $x^e=0$ *only if* (Σ) *has an asymptotically stable* Δ_{mix} *dynamics.*

Proof. Since Δ^e_{mix} is nonsingular, it is involutive and so has an integral manifold I^e passing through the equilibrium point $x^e=0$ of the drift term

$$\dot{x}^e=F(x^e)=\begin{pmatrix} f(x) \\ \bar{f}(x^e) \end{pmatrix}+\sum_{i=1}^{m}\begin{pmatrix} g_i(x) \\ 0 \end{pmatrix}\cdot\alpha_i(x^e) \tag{4.3}$$

of (Σ^e). Δ^e_{mix} is also invariant under F, and this implies that I^e, containing an equilibrium of (4.3), is an invariant manifold for (4.3). This proves claim (1). Consider now the Δ^e_{mix} dynamics, i.e., the restriction of (4.3) to I^e. If B is an arbitrary mixed bracket of (Σ^e), then, by the decoupling property, $L_B\alpha_i\equiv 0$, $i=1,\cdots,m$. Since our feedback preserves the equilibrium, $\alpha_i(0)=0$, so $\alpha_i\equiv 0$ on I^e. Similarly, using the invariant projection property for Δ^e_{mix}, we deduce that $x_1\equiv 0,\cdots,x_m\equiv 0$, $x^{**}_{m+1}\equiv 0$, $x_{m+2}\equiv 0$ on I^e, whereas the coordinate functions x^*_{m+1} can be made part of a coordinate system $(x^*_{m+1},\bar{x}^*)$ on I^e. Here $\bar{x}^*$ is an appropriate selection of components of $\bar{x}$. In these coordinates the restriction of (4.3) to I^e reads

$$\dot{x}^*_{m+1}=f^*_{m+1}(0,\cdots,0,x^*_{m+1},0,0),$$

$$\dot{\bar{x}}^*=\bar{F}(x^*_{m+1},\bar{x}^*),\quad \bar{F}\text{ appropriate.}$$

Claim (2) is now obvious. □

Remark 4.5. The mixed brackets distribution and the corresponding subsystem are structural properties typical for nonlinear systems. In fact, for a linear system all mixed brackets vanish and so the obstructive Δ_{mix} dynamics does not appear. From the general viewpoint of nonlinear control theory, the success of the linear Wonham-Morse approach thus obtains an additional motivation.

Remark 4.6. If the assumption of nonsingularity of Δ_{mix}, Δ^e_{mix} in Theorem 4.4 is dropped, we nevertheless can show the invariance of certain dynamical characteristics. In fact, since $f(0)=0$, $F(0)=0$, we can conclude by the mere definitions and without invoking nonsingularity that the subspaces $\Delta_{\text{mix}}(0)$, $\Delta^e_{\text{mix}}(0)$ are invariant subspaces, respectively, for the differentials $f_x(0)$ and $F_{x^e}(0)$ which correspond to the linearized drift terms of (Σ) and (Σ^e). What we can show is the following inclusion for the spectra of the respective restrictions:

$$\sigma(f_x(0)|_{\Delta_{\text{mix}}(0)}) \subset \sigma(F_{x^e}(0)|_{\Delta^e_{\text{mix}}(0)}). \tag{4.4}$$

In particular, (Σ^e) is surely unstable if an eigenvalue of $f_x(0)|_{\Delta_{\text{mix}}(0)}$ has a positive real part.

Proof. We have $\alpha_i(0)=0$, and by the decoupling property the restrictions of the differentials $d\alpha_i(0)$ to $\Delta^e_{\text{mix}}(0)$ also vanish. It follows that the restrictions on $\Delta^e_{\text{mix}}(0)$ of the linear maps given by

$$F_{x^e}(0) \quad \text{and} \quad \frac{d}{dx^e}\bigg|_0 \begin{pmatrix} f \\ \bar{f} \end{pmatrix} = \begin{pmatrix} f_x(0) & 0 \\ * & * \end{pmatrix}$$

coincide (cf. (4.3)). Formula (4.4) follows from this by using the invariant projection property of Δ^e_{mix} and some standard arguments from linear algebra. □

This remark and proof constitute a direct generalization of the concluding reasoning in [6, § 5].

5. Proof of Theorem 4.1. The theorem is proved once we succeed in showing that the projection π_x maps all members of a set of generators of Δ^e_{mix} into Δ_{mix} and that the images thus obtained suffice to generate all of Δ_{mix}. So our first step is the selection of appropriate sets of generators. Consider first Δ_{mix}. Let $L_{r,s} \subset \Delta_{\text{mix}}$ be the $\mathbb{R}$-linear span of all those mixed brackets which contain exactly r factors f and s factors in all ($0 \leqq r \leqq s-2$). By repeated use of the skew symmetry of the Lie product and the Jacobi identity, we show successively that $L_{r,s}$ is linearly spanned by each of the following sets of more special mixed brackets:

(i) The mixed brackets in $L_{r,s}$ of the form

$$[b_s, [b_{s-1}, \cdots [b_2, b_1] \cdots]],$$
$$b_s, \cdots, b_1 \in \{f, g_1, \cdots, g_m\}.$$

(ii) The mixed brackets in $L_{r,s}$ of the form

$$\begin{gathered}[(\text{ad}^{r_q}f, g_{i_q}), [\cdots, [(\text{ad}^{r_2}f, g_{i_2}), (\text{ad}^{r_1}f, g_{i_1})] \cdots]], \\ q=s-r, \qquad r_q+\cdots+r_1=r, \qquad 1 \leqq i_1, \cdots, i_q \leqq m\end{gathered} \tag{5.1}$$

(to obtain from (i) by distributing all factors f, beginning from the left, through the products on their right via the Jacobi identity. The ad notation is defined by $(\text{ad}^0 f, g)=g$, $(\text{ad}^{k+1}f, g)=[f, (\text{ad}^k f, g)]$). Note that $i_j \neq i_1$ for some j.

(iii) $E_{r,s}$ = the set of the mixed brackets in $L_{r,s}$ of the form (5.1) with the additional property $i_2 \neq i_1$.

To see this, use (ii) and the fact that a formal Lie bracket $[B, [A_1, \cdots, [A_{q-1}, A_q] \cdots]]$, where $B, A_1, \cdots, A_q$ are indeterminates, can be linearly combined by the brackets $[C_1, [\cdots, [C_q, B] \cdots]]$, where the C's form permutations of the A's. (This fact can easily be proved by induction on q.)

Thus, Δ_{mix} is generated by the union of the sets $E_{r,s}$. The corresponding quantities of the system (Σ^e) are defined in an analogous way and, as usual, marked by the superscript e. Now let

$$\tau_i = \sigma_i - \rho_i, \qquad i = 1, \cdots, m,$$

where σ_i (respectively, ρ_i) is the characteristic number of the output h_i with respect to (Σ^e) (respectively, (Σ)). Recall from § 2 that, if $\tau_i = 0$, then $\beta_{ii}(x^e) \neq 0$ near $x^e = 0$; otherwise, $\beta_{ii}(x^e) \equiv 0$ near $x^e = 0$ and $\tau_i - 1$ is just the characteristic number of $\alpha_i(x^e)$ with respect to (Σ^e).

LEMMA 5.1. *For $i = 1, \cdots, m$ and $\rho = 0, 1, 2, \cdots$ an identity of the following form holds*:

$$(5.2) \qquad \pi_x((\operatorname{ad}^\rho F, G_i)) = \sum_k b^\rho_{k,i} \cdot \phi^\rho_{k,i}.$$

Here

$b^\rho_{k,i}$ = *product of the form*:

$$(5.3) \qquad \begin{aligned} &[g_{j_r}, [\cdots, [g_{j_1}, g_i] \cdots]], \qquad 0 \leqq r \leqq \rho - \tau_i \ , \\ &j_r, \cdots, j_1 \in \{0, \cdots, m\} \quad (\text{here we let } g_0 = f); \end{aligned}$$

$\phi^\rho_{k,i}$ = *"product" of one of the following two types*:

$$D_r(\cdots (D_1(\beta_{ii})) \cdots), \qquad 0 \leqq r \leqq \rho,$$

$$D_r(\cdots (D_1(-L_{(\operatorname{ad}^t F, G_i)}\alpha_i)) \cdots), \qquad 0 \leqq r + t < \rho,$$

where $D_r, \cdots, D_1 \in \{L_F, \alpha_1, \cdots, \alpha_m\}$.

In addition,

(i) *For each k we have the following alternatives*:

—*Either $\phi^\rho_{k,i}$ contains no α_j, $j \neq i$, and therefore has the same decoupling property as α_i, β_{ii} themselves,*

—*Or $b^\rho_{k,i}$ is mixed, i.e., $b^\rho_{k,i}$ contains a factor g_j, $j \neq 0$, i.*

(ii) *If $\rho < \tau_i$, then* (5.2) *vanishes. Otherwise, exactly one of the $b^\rho_{k,i}$, say $b^\rho_{k^*,i}$, contains a maximal number of factors f; namely, we have*

$$b^\rho_{k^*,i} \cdot \phi_{k^*,i} = \begin{cases} (\operatorname{ad}^{\rho-\tau_i} f, g_i) \cdot \beta_{ii} & \text{if } \tau_i = 0, \\ -(\operatorname{ad}^{\rho-\tau_i} f, g_i) \cdot L_{(\operatorname{ad}^{\tau_i - 1} F, G_i)}\alpha_i & \text{if } \tau_i > 0 \end{cases}$$

and here $\phi_{k^*,i}(x^e) \neq 0$ near $x^e = 0$.

Remark. $\phi_{k^*,i}$ does not depend on ρ; therefore, the superscript ρ is dropped here. Note also that if $\tau_i > 0$ we have by the definition of the characteristic number $\tau_i - 1$ of α_i

$$(5.4) \qquad L_{(\operatorname{ad}^\rho F, G_i)}\alpha_i = \begin{cases} 0 & \text{for } 0 \leqq \rho < \tau_i - 1, \\ (-1)^{\tau_i - 1} L_{G_i} L_F^{\tau_i - 1}\alpha_i \neq 0 & \text{for } \rho = \tau_i - 1. \end{cases}$$

Proof of Lemma 5.1. Fix $i \in \{1, \cdots, m\}$. The proof proceeds by induction on ρ. For $\rho = 0$ we have by (2.6)

$$\pi_x(G_i) = g_i \cdot \beta_{ii} \quad (= 0 \text{ if } \tau_i > 0)$$

as desired. Induction step $\rho \to \rho + 1$: by assumption and (2.3a) there are vector fields $\bar{X}$, $\bar{X}' \in \operatorname{sp}\{\partial/\partial \bar{x}\}$ such that

$$(5.5) \qquad F = f + \sum_{j=1}^m g_j \alpha_j + \bar{X}, \qquad (\operatorname{ad}^\rho F, G_i) = \sum_k b^\rho_{k,i} \cdot \phi^\rho_{k,i} + \bar{X}'$$

(for notational convenience we write here f, g_j instead of $\binom{f}{0}$, $\binom{g_j}{0}$)). Now remember the following product rule for vector fields X_i, Y_j and scalar functions γ_i, δ_j:

$$X=\sum_i X_i\cdot\gamma_i,\qquad Y=\sum_j Y_j\cdot\delta_j\ \Rightarrow$$

$$[X,Y]=\sum_j Y_j\cdot(L_X\delta_j)-\sum_i X_i\cdot(L_Y\gamma_i)+\sum_{i,j}[X_i,Y_j]\cdot\gamma_i\delta_j.$$

Thus, bracketing the vector fields (5.5) we obtain

$$\begin{aligned}(\mathrm{ad}^{\rho+1}F,G_i)&=\sum_k b^{\rho}_{k,i}\cdot L_F\phi^{\rho}_{k,i}-\sum_{j=1}^{m}g_j\cdot L_{(\mathrm{ad}^{\rho}F,G_i)}\alpha_j\\ &\quad+\sum_k[f,b^{\rho}_{k,i}]\cdot\phi^{\rho}_{k,i}+[f,\bar{X}']+\sum_k\sum_{j=1}^{m}[g_j,b^{\rho}_{k,i}]\cdot\alpha_j\phi^{\rho}_{k,i}\\ &\quad+\sum_{j=1}^{m}[g_j,\bar{X}']\cdot\alpha_j+\sum_k[\bar{X},b^{\rho}_{k,i}]\cdot\phi^{\rho}_{k,i}+[\bar{X},\bar{X}'].\end{aligned}\tag{5.6}$$

Because of the decoupling property we have here

$$L_{(\mathrm{ad}^{\rho}F,G_i)}\alpha_j=0\quad\text{for } j\neq i.$$

In addition, sp$\{\partial/\partial\bar{x}\}$ is, of course, involutive and (trivially) invariant under $f, g_1,\cdots,g_m$, even under each $b\in\mathrm{Lie}\{f,\cdots,g_m\}$. It follows that all those terms in (5.6) that contain $\bar{X}$ or $\bar{X}'$ belong to sp$\{\partial/\partial\bar{x}\}$. Summing up, the action of π_x on (5.6) results exactly in a representation (5.2) with ρ now replaced by $\rho+1$. The additional statements on (5.2) follows by direct inspection of (5.6) using the induction hypothesis and (5.4). □

LEMMA 5.2. *Let* $0\leqq r\leqq s-2$, $q:=s-r$, $i_1,\cdots,i_q\in\{1,\cdots,m\}$, $i_1\neq i_2$, $r_q,\cdots,r_1$ *nonnegative integers such that* $r_q+\cdots+r_1=r$, $R_j:=r_j+\tau_{i_j}$ *for* $j=1,\cdots,q$. *Claim*:

$$\begin{aligned}&\pi_x([(\mathrm{ad}^{R_q}F,G_{i_q}),[\cdots,[(\mathrm{ad}^{R_2}F,G_{i_2}),(\mathrm{ad}^{R_1}F,G_{i_1})]\cdots]])\\ &\quad=[(\mathrm{ad}^{r_q}f,g_{i_q}),[\cdots,[(\mathrm{ad}^{r_2}f,g_{i_2}),(\mathrm{ad}^{r_1}f,g_{i_1})]\cdots]]\\ &\qquad\cdot\phi_{k^*,i_q}\cdots\phi_{k^*,i_1}+\sum_j b_j\cdot\psi_j(x^e),\end{aligned}\tag{5.7}$$

where

$$b_j\in\bigcup_{\substack{r'<r\\ s'\leqq s}}E_{r',s'}\cup\bigcup_{\substack{r'\leqq r\\ s'<s}}E_{r',s'},\quad \psi_j(x^e)\ \textit{smooth}$$

(*note that the first summand on the right-hand side is a nonzero multiple of a typical element of* $E_{r,s}$).

Proof. The proof is by induction on q. First, let $q=2$. For notational convenience we may assume without loss of generality that $i_1=1$, $i_2=2$. From Lemma 5.1 we get, for $i=1,2$,

$$(\mathrm{ad}^{R_i}F,G_i)=(\mathrm{ad}^{r_i}f,g_i)\cdot\phi_{k^*,i}+\sum_{k\neq k^*}b^{R_i}_{k,i}\cdot\phi^{R_i}_{k,i}+\bar{X}_i\tag{5.8}$$

with certain vector fields $\bar{X}_i\in$ sp$\{\partial/\partial\bar{x}\}$ and all brackets $b^{R_i}_{k,i}$ of the form (5.3). Since $k\neq k^*$, these brackets contain fewer than r_i factors f and, besides the obligatory factor

g_i, at most r_i more factors in all. We now bracket the vector fields (5.8) and obtain, by means of a computation similar to that in the preceding proof,

$$
\begin{aligned}
&[(\mathrm{ad}^{R_2}F, G_2), (\mathrm{ad}^{R_1}F, G_1)] \\
&\quad \in \sum_k b_{k,1}^{R_1} \cdot L_{(\mathrm{ad}^{R_2}F, G_2)}\phi_{k,1}^{R_1} - \sum_k b_{k,2}^{R_2} \cdot L_{(\mathrm{ad}^{R_1}F, G_1)}\phi_{k,2}^{R_2} \\
&\qquad + [(\mathrm{ad}^{r_2}f, g_2), (\mathrm{ad}^{r_1}f, g_1)] \cdot \phi_{k^*,2}\phi_{k^*,1} \\
&\qquad + \sum_{(k,k')\neq(k^*,k^*)} [b_{k,2}^{R_2}, b_{k',1}^{R_1}] \cdot \phi_{k,2}^{R_2}\phi_{k',1}^{R_1} + \mathrm{sp}\left\{\frac{\partial}{\partial \bar{x}}\right\}.
\end{aligned}
\tag{5.9}
$$

Here, in the first two sums, if a bracket b is not mixed then its ϕ factor has the decoupling property (see Lemma 5.1), and therefore the Lie derivative of this ϕ appearing in (5.9) vanishes. So the π_x projection of (5.9) contains only mixed brackets. The special form (5.7) is now verified just by counting the factors of each bracket.

The induction step $q \to q+1$ proceeds in virtually the same way as the initial step just completed. We need only the additional observation that the argument of π_x in (5.7), denoted for the moment by B, is now a mixed bracket. Therefore, by the decoupling property, $L_B\bar{\phi}_{k,i}^{\rho}=0$ for *all* functions ϕ in (5.2) (where $i=i_{q+1}$, $\rho=R_{q+1}$). It is for this reason that the induction step does not give rise to nonmixed brackets on the right-hand side of the new (5.7) corresponding to $q+1$ instead of q. □

LEMMA 5.3. *Let* $0 \leqq r \leqq s-2$. *Then* $\pi_x(E_{r,s}^e) \subset \Delta_{\mathrm{mix}}$.

Proof. For those members (5.1) (replace f, g by F, G) of $E_{r,s}^e$ whose exponents $r_q, \cdots, r_1$ are all sufficiently large, i.e., $r_j \geqq \tau_{i_j}$ for $j=1, \cdots, q$, the statement is contained in Lemma 5.2. For the others, the procedure of the proof of Lemma 5.2 can be followed all the same. The only difference is that now some of the sums (5.2) used in this proof may be empty. On the other hand, there is now no more need to identify certain nonzero contributions, so this does not matter. □

Theorem 4.1 now follows easily from these lemmas. In fact, by Lemma 5.3, $\pi_x(\Delta_{\mathrm{mix}}^e)=\sum_{r,s}\pi_x(\mathrm{sp}\, E_{r,s}^e) \subset \Delta_{\mathrm{mix}}$. On the other hand, it follows from Lemma 5.2 by a combined induction on (r, s) that $\pi_x(\Delta_{\mathrm{mix}}^e)$ indeed contains each $E_{r,s}$; therefore,

$$\pi_x(\Delta_{\mathrm{mix}}^e) \supset \mathrm{sp}\left(\bigcup_{r,s} E_{r,s}\right) = \Delta_{\mathrm{mix}}.$$

6. Examples.

Example 6.1. Isidori and Grizzle [6] proved that a system of the form

$$
\begin{aligned}
\dot{x}_1 &= u_1, & y_1 &= x_1, \\
\dot{x}_2 &= u_2, & y_2 &= x_2, \\
\dot{x}_3 &= a_3(x_1, x_2, x_3), & &
\end{aligned}
\tag{6.1}
$$

where

$$
\begin{aligned}
&a_3(0)=0, \\
&(a_3)_{x_1}(0)\neq 0, \quad (a_3)_{x_2}(0)\neq 0, \quad (a_3)_{x_3}(0)>0, \\
&(a_3)_{x_1x_2}(0)\neq 0
\end{aligned}
$$

cannot be rendered both noninteractive and stable, although it is already noninteractive and could also easily be stablized by a simple linear feedback. This fact is now easily explained by the theory just presented. The system (6.1) is already in P^* decomposed form and we have

$$P_1^* = \mathrm{sp}\left\{\frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}\right\}, \quad P_2^* = \mathrm{sp}\left\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_3}\right\}, \quad P^* = \mathrm{sp}\left\{\frac{\partial}{\partial x_3}\right\}.$$

Since

$$[g_2, [g_1, f]] = \begin{pmatrix} 0 \\ 0 \\ (a_3)_{x_1 x_2} \end{pmatrix} \in \Delta_{\mathrm{mix}},$$

we have here $\Delta_{\mathrm{mix}} = P^*$. So the Δ_{mix} dynamics equals the P^* dynamics

$$\dot{x}_3 = a_3(0, 0, x_3),$$

which is unstable because of $(a_3)_{x_3}(0) > 0$. The desired result now follows from our Remark 4.6. Its proof in [6] can be directly recognized as a special case of our preceding calculations.

Example 6.2. The next example demonstrates that a nonlinear system with an unstable P^* dynamics may nevertheless be rendered both noninteractive and stable if its Δ_{mix} dynamics is stable. Consider the following modification of the preceding example:

$$\text{(6.2)} \qquad \begin{aligned} &\dot{x}_1 = u_1, \qquad y_1 = x_1, \\ &\dot{x}_2 = u_2, \qquad y_2 = x_2, \\ &\dot{x}_3 = (x_1 + x_2)^2 - x_3, \\ &\dot{x}_4 = x_1 - x_2 + x_4. \end{aligned}$$

Obviously the linearization of (6.2), and hence (6.2) itself, are stabilizable at $x = 0$. Using (3.3) we compute

$$P_1^* = \mathrm{sp}\left\{\frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}, \frac{\partial}{\partial x_4}\right\},$$

$$P_2^* = \mathrm{sp}\left\{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_3}, \frac{\partial}{\partial x_4}\right\},$$

$$P^* = \mathrm{sp}\left\{\frac{\partial}{\partial x_3}, \frac{\partial}{\partial x_4}\right\}, \qquad \Delta_{\mathrm{mix}} = \mathrm{sp}\left\{\frac{\partial}{\partial x_3}\right\}.$$

Thus, (6.2) is in refined P^* decomposed form. Its P^* dynamics

$$\dot{x}_3 = -x_3, \qquad \dot{x}_4 = x_4$$

is obviously unstable, whereas the Δ_{mix} dynamics

$$\dot{x}_3 = -x_3$$

is asymptotically stable. So no static feedback can render (6.2) noninteractive and stable; yet a suitable dynamic feedback may exist, and it does indeed. Adjoin to (6.2) the equation

$$\dot{x}_5 = -x_2 + x_5.$$

In the new coordinates

$$w_1 = x_1, \qquad w_3 = x_2,$$
$$w_2 = x_4 - x_5, \qquad w_4 = x_5,$$

the (x_1, x_2, x_4, x_5) part of the thus extended system reads

$$\begin{aligned} \dot{w}_1 &= u_1, \qquad y_1 = w_1, \\ \dot{w}_2 &= w_1 + w_2, \\ \dot{w}_3 &= u_2, \qquad y_2 = w_3, \\ \dot{w}_4 &= -w_3 + w_4. \end{aligned}$$

This pair of independent controllable single-input single-output systems can, of course, easily be stabilized without affecting the noninteraction. The final adjunction of the x_3 part

$$\dot{x}_3 = (w_1 + w_2)^2 - x_3$$

obviously does not change matters.

In view of this example it is natural to ask whether the stability of the Δ_{mix} dynamics is also sufficient for noninteraction with stability in our setup. However, up to now this question is unsettled. The above procedure is not general. For instance, changing x_1 to x_1^3 in the $\dot{x}_4$ equation of (6.2) does not affect P^* and Δ_{mix}, but the procedure now leads to a (w_1, w_2) subsystem which is no longer stabilizable. It is unknown to the author whether this modified system can at all be rendered noninteractive and stable. What can be done in general is to construct dynamic noninteraction feedbacks which reduce the distribution P^{*e} of the resulting system (Σ^e) to the smallest possible dimension, namely, that of Δ_{mix}. Hence, there is no fixed dynamics bigger than the Δ_{mix} dynamics. Unfortunately, the feedbacks known up to now generally destroy the stabilizability (even in the linear case), so they are of little practical interest, and we forego a more detailed discussion.

Acknowledgments. The author thanks A. Isidori for many fruitful and encouraging discussions on the subject. The financial support of the Deutsche Forschungsgemeinschaft and of the University of Rome is also gratefully acknowledged.

REFERENCES

[1] J. Descusse and C. H. Moog, *Decoupling with dynamic compensation for strong invertible affine nonlinear systems*, Internat. J. Control, 42 (1985), pp. 1387–1398.

[2] M. Fliess, *A new approach to the noninteracting control problem in nonlinear system theory*, Proc. Allerton Conference on Commun. Contr. Comput., 1985, pp. 123–129.

[3] E. G. Gilbert, *The decoupling of multivariable systems by state feedback*, SIAM J. Control Optim., 7 (1969), pp. 50–63.

[4] I. J. Ha and E. G. Gilbert, *A complete characterization of decoupling control laws for a general class of nonlinear systems*, IEEE Trans. Automat. Control, 31 (1986), pp. 823–830.

[5] A. Isidori, *Nonlinear Control Systems: an Introduction*, Lecture Notes in Control and Information Sciences, Vol. 72, Springer-Verlag, Berlin, 1985.

[6] A. Isidori and J. W. Grizzle, *Fixed modes and nonlinear noninteracting control with stability*, IEEE Trans. Automat. Control, 33 (1988), pp. 907–914.

[7] H. Nijmeijer and W. Respondek, *Decoupling via dynamic compensation for nonlinear control systems*, in Proc. 25th Annual IEEE Conference on Decision and Control, IEEE Computer Society, Washington, DC, 1986, pp. 192–197.

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

[9] W. M. WONHAM, *Linear Multivariable Control: a Geometric Approach*, 3rd ed., Springer-Verlag, New York, 1985.

[10] W. M. WONHAM AND A. S. MORSE, *Decoupling and pole assignment in linear multivariable systems: a geometric approach*, SIAM J. Control Optim., 8 (1970), pp. 1–18.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 3, pp. 623-637, May 1991

007

OPTIMAL DESIGN OF DOMAINS WITH FREE-BOUNDARY PROBLEMS*

VIOREL BARBU† AND AVNER FRIEDMAN‡

Abstract. A domain $\tilde{\Omega}_u$ depends upon a control variable u, and its boundary consists of three parts: Γ_0, Γ_u, and $\tilde{\Gamma}_u$. The part Γ_0 is prescribed independently of u. The part Γ_u is determined directly upon prescribing the control u. The remaining part $\tilde{\Gamma}_u$ is determined as the free boundary of a variational inequality with $\tilde{\Omega}_u$ as the noncoincidence set. The problem is to choose u so that $\tilde{\Gamma}_u$ will become "as close as possible" to a given surface. Properties of the optimal control u^* are derived.

Key words. optimal control, variational inequalities, optimal design

AMS(MOS) subject classifications. 49A15, 49A35, 49A35, 34H05

1. The physical problem. Let Q be a bounded domain in $\mathbb{R}^n$ with C^2 boundary and let a, b, H be constants satisfying $0 \leqq a < b < H$. Let

$$Q_H = \{(x, y);\ x \in Q, 0 < y < H\}.$$

We introduce the set U of control functions $u(x)$:

$$U = \{u \in \text{Lip}\,(\bar{Q}), |\nabla u| \leqq \rho \text{ a.e., } a \leqq u(x) \leqq b, \text{ and } u = u_0 \text{ on } \partial Q\},$$

where u_0 is a given C^1 function satisfying

$$a < u_0 < b,$$

and ρ is a given positive number; U is a compact subset in $L^2(Q)$.

For any $u \in U$ let

$$\Omega_u = \{(x, y);\ x \in Q, u(x) < y < H\}$$

and consider the variational inequality

$$\int_{\Omega_u} \nabla z \cdot \nabla(\zeta - z)\,dx\,dy \geqq -\int_{\Omega_u} f(\zeta - z)\,dx\,dy \quad \forall \zeta \in K_u, \quad z \in K_u, \tag{1.1}$$

where

$$K_u = \{z \in H^1(\Omega_u),\ z \geqq 0 \text{ in } \Omega_u,\ z = 1 \text{ on } \{y = u(x)\},\ z = 0 \text{ on } \{y = H\}\} \tag{1.2}$$

and f is a given function in $C^{0,1}(\bar{Q}_H)$ satisfying

$$0 \leqq f \leqq C^*, \quad f_y \geqq 0 \quad (C_* \text{ positive constant}); \tag{1.3}$$

the last two conditions in (1.2) are taken in the usual trace sense (recall [12, § 4.2]).

Formally, problem (1.1) is equivalent to the obstacle problem

$$\begin{aligned} &-\Delta z = -f \quad \text{in } \{z > 0\}, \quad z \geqq 0 \quad \text{in } \Omega_u, \quad -\Delta z \geqq -f \quad \text{in } \Omega_u, \\ &z = 1 \quad \text{on } \{y = u(x)\}, \qquad z = 0 \quad \text{on } \{y = H\}, \\ &\frac{\partial z}{\partial \nu} = 0 \quad \text{on the lateral boundary of } \Omega_u. \end{aligned} \tag{1.1a}$$

* Received by the editors May 5, 1989; accepted for publication (in revised form) May 29, 1989. This work was partially supported by National Science Foundation grant DMS-86-12880.

† University of Iasi, Institute of Mathematics, Iasi, Romania.

‡ Institute for Mathematics and its Applications, University of Minnesota, Minneapolis, Minnesota 55455.

Let $y=\psi(x)$ be a given function satisfying

$$\psi\in C^2(\bar Q),\quad b<\psi(x)\leqq H\quad \text{if } x\in\bar Q. \tag{1.4}$$

We wish to find a control u such that the free boundary of the corresponding solution $z=z_u$ of (1.1), (1.2) will be "as close as possible" to $\{y=\psi\}$. This problem arises in electrochemical machinery [4, p. 177], [11] whereby we wish to achieve a specific shape of the surface of a metal workpiece (the anode) by electrochemical process; the control $\{y=u(x)\}$ plays the role of the cathode part of the outer boundary of the bath containing a chemical solution, and a fixed voltage is applied along this boundary. The function f represents a source term which we can introduce into the process.

In this paper we will adopt a weak formulation of the goal of achieving a shape close to $\{y=\psi(x)\}$ (see Remark 5.2), and we will establish necessary conditions on the optimal control u. In particular, the optimal control will be uniquely and explicitly determined for $u_0\equiv$ const., u_0 depending on the domain (see Corollary 5.2). At the end of the paper we will briefly consider other control sets.

2. The ε-problem. Let $X=(x,y)$, $dX=dx\,dy$.

Let $F(X,z)$ be a function satisfying

$$F\in C^1(\bar Q_H\times\mathbb{R}),\quad F\geqq -C_0\quad (C_0 \text{ const.}). \tag{2.1}$$

Introduce the functional

$$J(u)=\int_{\Omega_u}F(x,z_u)\,dX, \tag{2.2}$$

where z_u is the solution of (1.1), (1.2), and consider the following problem.

PROBLEM (P). Find u^* such that

$$u^*\in U,\quad J(u^*)=\min_{u\in U}J(u).$$

THEOREM 2.1. *There exists a solution u^* of Problem (P).*

Proof. Take a minimizing sequence u_m and denote the corresponding solution of (1.1), (1.2) by z_m. By standard estimates for variational inequalities [5], [9] we deduce that, for a subsequence, $u_m\to u^*$, $z_m\to z^*$ uniformly, and (u^*,z^*) is a solution of (1.1), (1.2). Clearly, u^* is then a solution of Problem (P).

From now on we will deal with a specific minimizer u^* and denote the corresponding solution of (1.1), (1.2) by z^*. We also set

$$\Omega^*=\Omega_{u^*}.$$

In order to derive necessary conditions on u^*, we wish to consider ε-approximate problems for which the cost function is differentiable; this procedure was used in [2] and [6].

Let $j\in C_0^\infty(\{|x|<1\})$, $j\geqq 0$, $\int j(x)\,dx=1$, $j_\varepsilon(x)=\varepsilon^{-n}j(x/\varepsilon)$ for any $\varepsilon>0$, and define the mollifier

$$\mu_\varepsilon(u)=\int_Q j_\varepsilon(x-\xi)u(\xi)\,d\xi.$$

Note that $\mu_\varepsilon(u)\to u$ in $C^\alpha(Q)$ for any $0<\alpha<1$.

Introduce functions $\beta_\varepsilon(z)\in C^\infty$ in z $(0<\varepsilon<1)$ such that

$$\begin{aligned}&\beta_\varepsilon(z)\to 0\quad \text{if } z>0,\quad \varepsilon\to 0,\qquad \beta_\varepsilon(z)\to-\infty\quad \text{if } z<0,\quad \varepsilon\to 0;\\ &\beta_\varepsilon'(z)\geqq 0\quad \text{for all } z,\quad \text{and}\quad \beta_\varepsilon(\varepsilon)=0,\quad \beta_\varepsilon(0)=-C_*\end{aligned} \tag{2.3}$$

with C_* as in (1.3). Consider the elliptic problem:

$$(2.4)\qquad -\Delta z+\beta_\varepsilon(z)=-f \quad \text{in } \Omega_{\mu_\varepsilon(u)},\quad 0\leqq\varepsilon<1,$$

$$z=1 \quad \text{on } \{y=\mu_\varepsilon(u)\},\qquad z=0 \quad \text{on } \{y=H\},$$

$$(2.5)\qquad \frac{\partial z}{\partial \nu}=0 \quad \text{on the lateral boundary of } \Omega_{\mu_\varepsilon(u)}.$$

We denote the solution by z_u^ε; note that z_u^ε belongs to $H^1(\Omega_{\mu_\varepsilon(u)})\cap C^{2,\alpha}(\Omega_{\mu_\varepsilon(u)})$. Consider the functional

$$J_\varepsilon(u)=\int_{\Omega_{\mu_\varepsilon(u)}} F(X,z_u^\varepsilon)\,dX+\frac{1}{2}\int_Q |u-u^*|^2\,dx.$$

PROBLEM (P_ε). Find $u_\varepsilon\in U$ such that

$$u_\varepsilon\in U,\qquad J_\varepsilon(u_\varepsilon)=\min_{u\in U} J_\varepsilon(u).$$

Proceeding as in the proof of Theorem 2.1 we can show that there exists a solution u_ε of Problem (P_ε); we denote the corresponding solution of (2.4), (2.5) by $z_{u_\varepsilon}^\varepsilon$. Furthermore, we have the following lemma.

LEMMA 2.2. *As* $\varepsilon\to 0$,

$$(2.6)\qquad u_\varepsilon\to u^* \quad \textit{weakly in } W^{1,p}(Q) \textit{ and strongly in } C^0(\bar{Q}),$$

$$(2.7)\qquad z_{u_\varepsilon}^\varepsilon\to z^* \quad \textit{weakly in } W^{2,p}_{\text{loc}}(\Omega_*)$$

for any $1<p<\infty$.

Proof. For any sequence $\varepsilon\to 0$ there is a subsequence such that

$$(2.8)\qquad u_\varepsilon\to u_1,\qquad z_{u_\varepsilon}^\varepsilon\to z_1$$

in the sense of (2.6), (2.7). Furthermore, taking $\varepsilon\to 0$ in the relation

$$J_\varepsilon(u_\varepsilon)\leqq J_\varepsilon(u^*),$$

we arrive at the inequality

$$\int_{\Omega_{u_1}} F(X,z_1)+\frac{1}{2}\limsup_{\varepsilon\to 0}\int_Q |u_\varepsilon-u^*|^2\leqq\int_{\Omega^*} F(X,z^*)\leqq J(u^*).$$

From the optimality of u^* it then follows that

$$\limsup_{\varepsilon\to 0}\int_Q |u_\varepsilon-u^*|^2=0$$

and therefore also $u_1=u^*$, $z_1=z^*$. Clearly, (2.6), (2.7) now follow from (2.8).

LEMMA 2.3. *For any* $u\in U$ *the corresponding solution* $z_\varepsilon\equiv z_{u_\varepsilon}^\varepsilon$ *of* (2.4), (2.5) *satisfies*

$$(2.9)\qquad 0\leqq z_\varepsilon\leqq 1,$$

$$(2.10)\qquad \frac{\partial z_\varepsilon}{\partial y}\leqq 0.$$

Proof. Suppose z_ε takes negative minimum at a point $X_0=(x_0,y_0)$ in $\Omega_{\mu_\varepsilon(u)}$. Then $\beta_\varepsilon(z_\varepsilon(X_0))<-C_*$ by (2.3). Since also $\Delta z_\varepsilon(X_0)\geqq 0$, (2.4) gives $f(X_0)>C_*$, a contradiction to (1.3). Thus the minimum of z_ε is achieved only on the boundary, and (by (2.5)) it is greater than or equal to zero.

Since $\beta_\varepsilon(1)=0$, $z\equiv 1$ is a supersolution of (2.4), (2.5); it follows (by comparison) that $z_\varepsilon \leqq 1$.

The function $w=\partial z_\varepsilon/\partial y$ satisfies

$$-\Delta w+\beta_\varepsilon'(z_\varepsilon)w=-f_y \quad \text{in } \Omega_{\mu_\varepsilon(u)},$$

and

$$\begin{aligned} &w\leqq 0 && \text{on } \{y=H\} \quad (\text{since } z_\varepsilon \geqq 0 \text{ in } \Omega_{\mu_\varepsilon(u)}),\\ &w\leqq 0 && \text{on } \{y=\mu_\varepsilon(u)\} \quad (\text{since } z_\varepsilon \leqq 1 \text{ in } \Omega_{\mu_\varepsilon(u)}),\\ &\frac{\partial w}{\partial \nu}=0 && \text{on the lateral boundary of } \Omega_{\mu_\varepsilon(u)}. \end{aligned}$$

Hence proceeding formally to use the maximum principle, we deduce that $w\leqq 0$, and (2.10) thus follows. Since w may not actually be continuous (or even bounded) at the corner points of $\Omega_{\mu_\varepsilon(u)}$, we should proceed to establish assertion (2.10) either by approximating $y=\mu_\varepsilon(u)$ so that it becomes horizontal near ∂Q (and then w is continuous in $\overline{\Omega_{\mu_\varepsilon(u)}}$) or else by working with the weak equation for w and applying the maximum principle for weak solutions.

3. The optimality conditions for (P_ε). We denote by u_ε a solution of Problem (P_ε) and by z_ε the corresponding solution of (1.1), (1.2); set $\Omega_\varepsilon=\Omega_{\mu_\varepsilon(u_\varepsilon)}$, $\Gamma=\{(x,H);\ x\in Q\}$, $\Gamma_\varepsilon=\{(x,y);\ y=\mu_\varepsilon(x), x\in Q\}$, $\Gamma_{1,\varepsilon}=$ the lateral boundary of Ω_ε.

Let v be any function such that $u_\varepsilon+\delta v\in U$ for all sufficiently small $\delta>0$, and introduce the quotient

$$q_{\varepsilon,\delta}=\frac{z^\varepsilon_{u_\varepsilon}+\delta v-z_\varepsilon}{\delta}.$$

LEMMA 3.1. *As* $\delta\to 0$,

$$q_{\varepsilon,\delta}\to w_\varepsilon \quad \textit{weakly in } W^{2,r}_{\text{loc}}(\Omega_\varepsilon) \tag{3.1}$$

for any $1<r<\infty$, *where* w_ε *is the weak solution in* $H^1(\Omega_\varepsilon)$ *of*

$$-\Delta w_\varepsilon+\beta_\varepsilon'(z_\varepsilon)w_\varepsilon=0 \quad \textit{in } \Omega_\varepsilon, \tag{3.2}$$

$$w_\varepsilon(x,y)=-\frac{\partial z_\varepsilon(x,y)}{\partial y}\mu_\varepsilon(v)(x) \quad \textit{on } \Gamma_\varepsilon, \tag{3.3}$$

$$w_\varepsilon=0 \quad \textit{on } \Gamma, \tag{3.4}$$

$$\frac{\partial w_\varepsilon}{\partial \nu}=0 \quad \textit{on the lateral boundary } \Gamma_{1,\varepsilon}. \tag{3.5}$$

Remark 3.1. Since w_ε is only in $H^1(\Omega_\varepsilon)$, it should be mentioned that $\partial w_\varepsilon/\partial\nu$ makes sense as an element of $H^{-1/2}(\Gamma_{1,\varepsilon})$, and the solution to (3.2)-(3.5) is defined in the sense that

$$\begin{aligned} &\int_{\Omega_\varepsilon} w_\varepsilon(x,y)[\zeta(x,y)\beta_\varepsilon'(z_\varepsilon)-\Delta\zeta(x,y)]\,dX\\ &\quad=\int_{\Gamma_\varepsilon}\frac{\partial z_\varepsilon(x,y)}{\partial y}\frac{\partial\zeta(x,y)}{\partial\nu}\mu_\varepsilon(v)(x)\,dS \end{aligned} \tag{3.6}$$

for all $\zeta\in H^2(\Omega_\varepsilon)$ such that $\zeta=0$ on $\Gamma\cup\Gamma_\varepsilon$ and $\partial\zeta/\partial\nu=0$ on $\Gamma_{1,\varepsilon}$.

Proof. Taking the difference of the elliptic equations for $z^\varepsilon_{u_\varepsilon+\delta v}$ and z_ε and applying L^p elliptic estimates, we find that

$$\|z^\varepsilon_{u_\varepsilon}+\delta v-z_\varepsilon\|_{W^{2,r}(\Omega^*)}\leqq\tilde{C}\delta$$

for any compact subdomain Ω^* of Ω_ε; $\tilde{C}$ depends on ε and Ω^*. It follows that for any sequence $\delta \to 0$ there is a subsequence for which (3.1) holds. Since (3.4) and (3.5) are rather obviously satisfied, it remains to establish (3.3). Using the fact that the boundaries $\{y = \mu_\varepsilon(u_\varepsilon)\}$ and $\{y = \mu_\varepsilon + \delta v)\}$ are smooth, we can compute $q_{\varepsilon,\delta}$ on $\{y = \mu_\varepsilon(u_\varepsilon) + C\delta\}$, for any $C > \sup v$. Setting $\theta_\delta(x) = \mu_\varepsilon(u_\varepsilon)(x) + C\delta$, we deduce that

$$q_{\varepsilon,\delta}(x, \theta_\delta(x)) = -\frac{\partial z^\varepsilon_{u_\varepsilon + \delta v}}{\partial y}(x, \theta_\delta(x))[\mu_\varepsilon(u_\varepsilon + \delta v)(x) - \theta_\delta(x)]$$

$$+\frac{\partial z_\varepsilon}{\partial y}(x, \theta_\delta(x))[\mu_\varepsilon(u_\varepsilon)(x) - \theta_\delta(x)] + o(1)$$

as $\delta \to 0$, from which (3.3), in the sense of (3.6), easily follows.

For any convex set U in a real Banach space X we define

$$h_U(u) = \begin{cases} 0 & \text{if } u \in U, \\ +\infty & \text{if } u \notin U. \end{cases}$$

Recall [1] that the subdifferential $\partial h_U(u)$ is defined as follows:

$w \in \partial h_U(u)$ if and only if $w \in X^*$ (the dual of X) and $h_U(u+v) - h_U(u) \geqq (w, v)$ $\forall v \in X$;

here (w, v) denotes the application of the bounded linear functional w to v. Recalling the definition of $h_U(u+v)$, we see that

$$\partial h_U(u) = \{w \in X;\ (w, u - v) \geqq 0\ \forall v \in U\}. \tag{3.7}$$

Since $u \to h_U(u)$ is a convex function, we easily find that

(3.8) $w \in \partial h_U(u)$ if and only if $\limsup_{\delta \downarrow 0} (h_U(u + \delta v) - h_U(u))/\delta \geqq (w, v)$, for all $v \in X$ such that $u + \delta v \in U$.

From the optimality of $(u_\varepsilon, z_\varepsilon)$ we have

$$\int_{\Omega_{\mu_\varepsilon}(u_\varepsilon + \delta v)} F(X, z^\varepsilon_{u_\varepsilon + \delta v)} - \int_{\Omega_\varepsilon} F(X, z_\varepsilon) + \int_Q \left[u_\varepsilon - u^* + \frac{\delta}{2} v\right] \delta v \geqq 0 \quad \text{if } u_\varepsilon + \delta v \in U. \tag{3.9}$$

Dividing by δ and letting $\delta \to 0$, we get

$$\begin{aligned} &\int_{\Omega_\varepsilon} F_z(X, z_\varepsilon) w_\varepsilon - \int_Q F(X, z_\varepsilon(x, \mu_\varepsilon(u_\varepsilon)(x))) \mu_\varepsilon(v)(x)\, dx \\ &\quad + \int_Q (u_\varepsilon - u^*) v \geqq 0 \quad \text{if } u_\varepsilon + \delta v \in U \quad \text{for all small } \delta > 0. \end{aligned} \tag{3.10}$$

We will rewrite this condition more effectively by introducing the "adjoint variable" p_ε, defined as the solution of the elliptic problem:

$$\begin{aligned} &-\Delta p_\varepsilon + \beta'_\varepsilon(z_\varepsilon) p_\varepsilon = -F_z(X, z_\varepsilon) \quad \text{in } \Omega_\varepsilon, \\ &p_\varepsilon = 0 \quad \text{on } \Gamma \cup \Gamma_\varepsilon, \\ &\frac{\partial p_\varepsilon}{\partial \nu} = 0 \quad \text{on } \Gamma_{1,\varepsilon}. \end{aligned} \tag{3.11}$$

Note that p_ε is smooth up the boundary $\{y = \mu_\varepsilon(u_\varepsilon)\}$; furthermore,

$$p_\varepsilon \in H^1(\Omega_\varepsilon) \cap H^2(\Omega^\delta_\varepsilon) \quad \forall \text{ small } \delta > 0,$$

where

$$\Omega_\varepsilon^\delta = \{(x, y);\ x \in Q, \operatorname{dist}(x, \partial Q) > \delta, \mu_\varepsilon(u_\varepsilon(x)) < y < H\}.$$

We can transform the first integral on the left-hand side of (3.10), formally as follows:

$$\begin{aligned}\int_{\Omega_\varepsilon} F_z(X, z_\varepsilon) w_\varepsilon &= \int_{\Omega_\varepsilon} [\Delta p_\varepsilon - \beta'_\varepsilon(z_\varepsilon) p_\varepsilon] w_\varepsilon \\ &= \int_{\partial\Omega_\varepsilon} \left(\frac{\partial p_\varepsilon}{\partial \nu} w_\varepsilon - p_\varepsilon \frac{\partial w_\varepsilon}{\partial \nu}\right) \quad \text{(by (3.2))} \\ &= -\int_Q \left[\frac{\partial z_\varepsilon}{\partial y} \cdot \frac{\partial p_\varepsilon}{\partial \nu}\right]_{y=\mu_\varepsilon(u_\varepsilon)} \mu_\varepsilon(v)[1+|\nabla \mu_\varepsilon(u_\varepsilon)|^2]^{1/2}\, dx \quad \text{(by (3.3)-(3.5))}.\end{aligned} \tag{3.12}$$

To proceed rigorously we take $\alpha_\delta \in C_0^\infty(\Omega_\varepsilon)$ such that $\alpha_\delta = 1$ in $\Omega_\varepsilon^\delta$ and $\alpha_\delta = 0$ in $\Omega_\varepsilon \backslash \Omega_\varepsilon^{\delta/2}$. We set $q_{\varepsilon,\delta} = p_\varepsilon \alpha_\delta$ and note that

$$q_{\varepsilon,\delta} \in H^2(\Omega_\varepsilon), \quad -\Delta q_{\varepsilon,\delta} + \beta'_\varepsilon(z_\varepsilon) q_{\varepsilon,\delta} = -\alpha_\delta F_z(X, z_\varepsilon) + \rho_{\varepsilon,\delta} \quad \text{in } \Omega_\varepsilon,$$

$$q_{\varepsilon,\delta} = 0 \quad \textit{on } \Gamma \cup \Gamma_\varepsilon, \qquad \frac{\partial q_{\varepsilon,\delta}}{\partial \nu} = 0 \quad \text{on } \Gamma_{1,\varepsilon},$$

where

$$\|\rho_{\varepsilon,\delta}\|_{L^2(\Omega_\varepsilon)} \leqq C\delta.$$

If we take $\zeta = q_{\varepsilon,\delta}$ in (3.6), we get the following precise (and modified) version of (3.12):

$$\begin{aligned}\int_{\Omega_\varepsilon} F_z(X, z_\varepsilon) w_\varepsilon = -\int_Q &\left[\frac{\partial z_\varepsilon}{\partial y} \cdot \frac{\partial q_{\varepsilon,\delta}}{\partial \nu}\right]_{y=\mu_\varepsilon(u_\varepsilon)} \mu_\varepsilon(v)[1+|\nabla \mu_\varepsilon(u_\varepsilon)|^2]^{1/2}\, dx \\ &+ \int_{\Omega_\varepsilon} [F_z(X, z_\varepsilon)(1-\alpha_\delta) - \rho_{\varepsilon,\delta}] w_\varepsilon.\end{aligned} \tag{3.13}$$

Substituting this into (3.10), we get

$$\begin{aligned}\int_Q \Bigg\{ -\mu_\varepsilon &\left\{\frac{\partial z_\varepsilon}{\partial y} \cdot \frac{\partial q_{\varepsilon,\delta}}{\partial \nu} [1+|\nabla \mu_\varepsilon(u_\varepsilon)|^2]^{1/2}\right\} \\ &- \mu_\varepsilon\{F(X, z_\varepsilon)\mu_\varepsilon(u_\varepsilon)\} + (u_\varepsilon - u^*) - \varphi_\delta \Bigg\} v \geqq 0\end{aligned} \tag{3.14}$$

if $u_\varepsilon + \delta v \in U$ for all small $\delta > 0$, where

$$\|\varphi_\delta\|_{L^2(\Omega)} \leqq C\delta.$$

Since $p_\varepsilon = q_{\varepsilon,\delta}$ in $\Omega_\varepsilon^\delta$, letting $\delta \to 0$ in (3.14) and recalling (3.8), we conclude that

$$\begin{aligned}\mu_\varepsilon &\left[\frac{\partial z_\varepsilon}{\partial y}(x, \mu_\varepsilon(u_\varepsilon)(x) \frac{\partial p_\varepsilon}{\partial \nu}(x, \mu_\varepsilon(u_\varepsilon)(x)(1+|\nabla \mu_\varepsilon(u_\varepsilon)|^2)^{1/2}\right] \\ &+ \mu_\varepsilon[F(X, z_\varepsilon(x, \mu_\varepsilon(u_\varepsilon)(x)))] \in \partial h_U(u_\varepsilon + (u_\varepsilon - u^*)),\end{aligned} \tag{3.15}$$

where U is the control set, considered as a subset of $L^2(Q)$. We summarize in Theorem 3.2.

THEOREM 3.2. *If* $(u_\varepsilon, z_\varepsilon, \Omega_\varepsilon)$ *is a solution of Problem* (P_ε), *then there exists a unique solution* p_ε *of* (3.11) *such that* (3.15) *holds.*

In order to apply this result we need to analyze the structure of $\partial h_U(u)$; this is done in the following section.

4. The structure of ∂h_U.

THEOREM 4.1. *Assume that there is a $g \in W^{1,\infty}(Q)$ such that $g = u_0$ in ∂Q and $\|\nabla g\|_{L^\infty(Q)} < \rho$. Then, for any $u \in U$, $w \in \partial h_U(u)$ is and only if w has the form*

$$w = -\text{div}\,\theta + \eta \quad \textit{in } Q, \tag{4.1}$$

where

$$\begin{aligned} &\theta \in (L^1(Q))^n, \\ &\theta(x) = 0 \quad \textit{a.e. in } \{x \in Q; |\nabla u(x)| < \rho\}, \\ &\theta(x) = \lambda(x)\nabla u(x) \quad \textit{a.e. in } \{x \in Q; |\nabla u(x)| = \rho\}, \\ &\textit{where } \lambda \in L^2(Q), \quad \lambda(x) \geqq 0 \quad \textit{a.e. } x \in Q \end{aligned} \tag{4.2}$$

and

$$\begin{aligned} &\eta \in L^2(Q), \\ &\eta(x) = 0 \quad \textit{a.e. in } \{x \in Q;\ a < u(x) < b\}, \\ &\eta(x) \leqq 0 \quad \textit{a.e. in } \{x \in Q;\ u(x) = a\}, \\ &\eta(x) \geqq 0 \quad \textit{a.e. in } \{x \in Q;\ u(x) = b\}. \end{aligned} \tag{4.3}$$

Proof. Observe first that if w is given by (4.1)-(4.3) then for any $v \in U$,

$$\begin{aligned} \int_Q w(x)(u(x) - v(x))\,dx &= -\int_Q \text{div}\,\theta(u - v)\,dx + \int_Q \eta(u - v)\,dx \\ &\geqq -\int_Q \text{div}\,\theta(u - v)\,dx = \int_Q (\theta, \nabla(u - v))\,dx \geqq 0; \end{aligned}$$

thus w belongs to $\partial h_U(u)$.

To prove the converse, write $U = U_0 \cap U_1$ where

$$U_0 = \{u \in W^{1,\infty}(Q); |\nabla u(x)| \leqq \rho \text{ a.e. in } Q;\ u = u_0 \text{ on } \partial Q\},$$
$$U_1 = \{u \in L^2(Q);\ a \leqq u(x) \leqq b \text{ a.e. in } \in Q\}$$

and set $h_i = h_{U_i}$, $i = 0, 1$.

It is easily seen that

$$\eta \in \partial h_1(u) \tag{4.4}$$

if and only if η satisfies (4.3).

Next we show that

$$w \in \partial h_0(u) \tag{4.5}$$

if and only if $w = -\text{div}\,\theta$ where θ satisfies (4.2). It is easily seen that if $w = -\text{div}\,\theta$ where θ satisfies (4.2) then w satisfies (4.5). To prove the converse, let $w \in \partial h_0(u) \subset L^2(Q)$. Since $w \in L^2(Q)$, there is $\xi \in (H^1(Q))^n$ such that $w(x) = -\text{div}\,\xi(x)$ almost everywhere in Q. By definition of $\partial h_0(u)$ we then find that $\xi \in \partial h_K(\nabla u)$ almost everywhere in Q where

$$K = \{\eta \in (L^\infty(Q))^n;\ \eta = \nabla v \text{ a.e. in } Q, v \in U_0\};$$

that is,

$$\int_Q (\xi, \nabla u - \eta)\,dx \geqq 0 \quad \forall \eta \in K. \tag{4.6}$$

We may represent K as $K=K_1\cap K_2$, where

$$K_1=\{\eta\in L^\infty(Q))^n;\ \eta=\nabla v,\ v\in W^{1,\infty}(Q),\ v=u_0 \text{ on } \partial Q\}.$$
$$K_2=\{\eta\in L^\infty(Q))^n;\ |\eta(x)|\leqq\rho \text{ a.e. in } Q\}.$$

Let us denote again $\partial h_K:(L^\infty(Q))^n\to((L^\infty))^n)^*$ the subdifferential of h_K viewed as function from $(L^\infty(Q))^n$ to $\bar{R}=(-\infty,+\infty]$. Then

$$\partial h_K(\eta)=\{\mu\in((L^\infty(Q))^n)^*;\ \mu(\eta-\nu)\geqq 0\ \forall\nu\in K\}$$

where $((L^\infty(Q))^n)^*$ is the dual space of $L^\infty(Q))^n$ and $\mu(\eta-\nu)$ is the value of measure μ in $\eta-\nu$. By assumption of the theorem we see that $g\in(\text{int } K_2)\cap K_1$ where int is taken in the $(L^\infty(Q))^n$ topology. Since $(\text{int } K_2)\cap K_1$ is nonempty, according to a well-known result due to Rockafellar [14] we have

$$\partial h_K=\partial h_{K_1}+\partial h_{K_2},$$

where $\partial h_{K_i}:(L^\infty(Q))^n\to((L^\infty(Q))^n)^*$ are subdifferentials of h_{K_i} viewed as functions from $(L^\infty(Q))^n$ into $\bar{R}$. Hence $\xi=\mu_1+\mu_2$ where $\mu_i\in\partial h_{K_i}(\nabla u)$. Let us denote by μ_i^a the absolutely continuous part of the measure μ_i. Then $\mu_i^a\in(L^1(Q))^n$ and, since $\xi\in(L^1(Q))^n$ (in fact ξ is even in $(H^1(Q))^n$), we must have that $\xi=\mu_1^a+\mu_2^a$.

We have

$$h_{K_2}(\eta)=\int_\Omega h(\eta(x))\,dx\quad\forall\eta\in(L^\infty(Q))^n,$$

where $h(\eta)=0$ if $|\eta|\leqq\rho$ and $h(\eta)=+\infty$ if $|\eta|>\rho$. But then, by Corollary 1.3 of [15],

$$\mu_2^a(x)\in\partial h(\nabla u(x))\quad\text{a.e. in } Q.$$

Since $\partial h(\eta)=0$ for $|\eta|<\rho$, $\partial h(\eta)=\{\lambda\eta,\ \lambda\geqq 0\}$ if $|\eta|=\rho$, we conclude that

$$\begin{aligned}&\mu_2^a(x)=0 &&\text{a.e. in } \{x\in Q;\ |\nabla u(x)|<\rho\},\\ &\mu_2^a(x)=\lambda\nabla u(x) &&\text{a.e. in } \{x\in Q;\ |\nabla u(x)|=\rho\},\end{aligned}\tag{4.7}$$

where $\lambda\in L^2(Q)$, $\lambda\geqq 0$ almost everywhere in Q.

Since $\partial h_{K_1}(\nabla u)$ is just the set of normals (in $((L^\infty(Q))^n)^*$) to the linear subspace $\{\eta\in(L^\infty(Q))^n;\ \eta=\nabla v,\ v\in W_0^{1,\infty}(Q)\}$ at ∇u, so is μ_1^a. Hence,

$$\int_Q \mu_1^a(x)\nabla v(x)\,dx=0\quad\forall v\in W_0^{1,\infty}(Q),\ v=0 \text{ on } \partial Q,$$

which yields

$$\operatorname{div}\mu_1^a=0\quad\text{a.e. in } Q.\tag{4.8}$$

Recalling that $\xi=\mu_1^a+\mu_2^a$ and using (4.7), (4.8) it follows that $w=-\operatorname{div}\xi=-\operatorname{div}\theta$ where $\theta=\xi-\mu_1^a$ satisfies (4.2).

To complete the proof of Theorem 4.1 we will need the following lemma [1, p. 28].

LEMMA 4.2. *Let A, B be maximal monotone operators in a real Hilbert space and assume that $D(A)\cap D(B)\neq 0$ and that*

$$(\xi,\eta)\geqq 0\quad\forall\xi\in Au,\quad\eta=B_\lambda u,\quad u\in D(A),\tag{4.9}$$

where $B_\lambda=\lambda^{-1}(1-(1+\lambda B^{-1}))$. Then $A+B$ is a maximal monotone operator.

We will apply this lemma with $A=\partial h_0$ and $B=\partial h_1$. We first observe that

$$(\xi,\eta)=0\quad\forall\xi\in Au,\quad\eta\in Bu,\quad u\in D(A)\cap D(B)\tag{4.10}$$

since by (4.4) and (4.5), $\xi=0$ almost everywhere in $\{x;\ \eta(x)\neq 0\}$. Next we note that $(1+\lambda B)^{-1}D(A)\subset D(A)$ for all $\lambda>0$. Indeed for any $u\in D(A)=U_0$,

$$(1+\lambda B)^{-1}u=\operatorname{Proj}\frac{u}{D(B)}=\begin{cases} u & \text{a.e. in } \{x;\ a\leqq u(x)\leqq b\},\\ a & \text{a.e. in } \{x;\ u(x)\leqq a\},\\ b & \text{a.e. in}\{x;\ u(x)\geqq b\}.\end{cases}$$

Thus since $\alpha\leqq u_0\leqq b$ the latter implies that $(1+\lambda B)^{-1}u\in U_0$. Also, by monotonicity of A,

$$(Au-A(1+\lambda B)^{-1}u,\ B_\lambda u)\geqq 0.$$

Since $B_\lambda u\in B(1+\lambda B)^{-1}u$, we conclude (using (4.10)) that

$$(Au,\ B_\lambda u)\geqq (A(1+\lambda B)^{-1}u,\ B(1+\lambda B)^{-1}u)\geqq 0$$

for any $u\in D(A)$.

We have thus verified the condition (4.9). It follows that the mapping $u\to \partial h_0(u)+\partial h_1(u)$ is maximal monotone. Since $\partial h_0+\partial h_1\subset \partial h_U$ we conclude that, in fact,

$$\partial h_0(u)+\partial h_1(u)=\partial h_U(u)$$

and Theorem 4.1 follows from the preceding characterizations of (4.4), (4.5).

5. Properties of the minimizer u^*. In this section we will apply Theorem 3.2 in order to deduce properties of the minimizer u^*. We take

$$F(X,z)=(z-z^0(X))^2, \tag{5.1}$$

where $z^0(x,y)$ satisfies

$$\Delta z^0\leqq f \quad \text{in } \bar{Q}_H, \tag{5.2}$$

$$z^0>1 \quad \text{in } \bar{Q}_H\cap\{a\leqq y\leqq b\}, \tag{5.3}$$

$$z^0>0 \quad \text{in } \overline{Q_H}, \tag{5.4}$$

$$\frac{\partial z^0}{\partial \nu}\geqq 0 \quad \text{on } \partial Q\times(0,H). \tag{5.5}$$

By comparison, $u^*<z^0$ in $\bar{Q}_H$ and, hence, if ε is sufficiently small,

$$z_\varepsilon-z^0\leqq -c_0<0 \quad \text{in } \Omega_\varepsilon, \tag{5.6}$$

where c_0 is a constant independent of ε. It follows that

$$-F_z(X,z_\varepsilon(X))=-2(z_\varepsilon-z^0)\geqq 2c_0>0 \quad \text{in } \Omega_\varepsilon.$$

Hence, we can apply the maximum principle to p_ε (in (3.11)) and deduce that

$$p_\varepsilon>0 \quad \text{in } \Omega_\varepsilon. \tag{5.7}$$

Since $p_\varepsilon=0$ on $\{y=\mu_\varepsilon(u_\varepsilon)\}$ we also get

$$\frac{\partial p_\varepsilon}{\partial \nu}<0 \quad \text{on } \{y=\mu_\varepsilon(u_\varepsilon)(x)\}. \tag{5.8}$$

Recalling (2.10) we conclude that the first term on the left-hand side of (3.15) is nonnegative. From (5.1), (5.6) we get

$$F(X, z_\varepsilon(X)) \geqq c_1 > 0 \quad \text{on } \Omega_\varepsilon,$$

where c_1 is a constant independent of ε. It follows that the left-hand side of (3.15) is strictly positive, independently of ε; i.e.,

$$w_\varepsilon + (u_\varepsilon - u^*) \geqq 2c > 0 \quad \text{for some } w_\varepsilon \in \partial h_U(u_\varepsilon), \tag{5.9}$$

where c is a constant independent of ε; in view of (2.6),

$$w_\varepsilon \geqq c > 0 \quad \text{for some } w_\varepsilon \in \partial h_U(u_\varepsilon) \tag{5.10}$$

if ε is sufficiently small. We will assume that

$$\|\nabla u_0\|_{L^\infty(Q)} < \rho. \tag{5.11}$$

Then, by Theorem 4.1 applied to w_ε,

$$-\text{div } \theta_\varepsilon + \eta_\varepsilon \geqq c > 0 \quad \text{a.e. in } Q, \tag{5.12}$$

where $\theta_\varepsilon \in (L^1(Q))^n$ and $\eta_\varepsilon \in L^2(Q)$ satisfy (4.2), (4.3); i.e.,

$$\nabla u_\varepsilon(x) = \rho \frac{\theta_\varepsilon(x)}{|\theta_\varepsilon(x)|} \quad \text{a.e. in } \{x \in Q;\ \theta_\varepsilon(x) \neq 0\}, \quad u_\varepsilon = u_0 \quad \text{on } \partial Q, \tag{5.13}$$

$$\begin{aligned} \eta_\varepsilon(x) &= 0 \quad \text{a.e. in } \{x;\ a < u_\varepsilon(x) < b\}, \\ \eta_\varepsilon(x) &\leqq 0 \quad \text{a.e. in } \{x;\ u_\varepsilon(x) = a\}, \\ \eta_\varepsilon(x) &\geqq 0 \quad \text{a.e. in } \{x;\ u_\varepsilon(x) = b\}. \end{aligned} \tag{5.14}$$

In particular, we see by (5.12) and (5.14) that $\theta_\varepsilon(x) \neq 0$ almost everywhere in $\{x \in Q;\ u_\varepsilon(x) < b\}$ and, therefore,

$$|\nabla u_\varepsilon(x)| = \rho \quad \text{a.e. in } Q_\varepsilon \equiv \{x \in Q;\ u_\varepsilon(x) < b\}. \tag{5.15}$$

In other words, $u_\varepsilon \in W^{1,\infty}(Q)$ is a weak solution in Q_ε of the eikonal equation

$$|\nabla u_\varepsilon| = \rho. \tag{5.16}$$

Since, in general, equation (5.16) with Dirichlet condition has an infinite number of weak solutions (i.e., solutions $u \in W^{1,\infty}(Q_\varepsilon)$ that satisfy almost everywhere the equation) it is not clear for the moment what u_ε looks like.

Recall [3] that $u \in C(G)$ (G an open subset of $\mathbb{R}^n$) is called a viscosity solution of (5.16) provided, for any $\phi \in C^1(G)$, if $u - \phi$ attains a local maximum at $x_0 \in G$ then $|\nabla \phi(x_0)| \leqq \rho$, and if $u - \phi$ attains a local minimum at $x_0 \in G$ then $|\nabla \phi(x_0)| \geqq \rho$.

An equivalent definition is the following (see [3]): Set

$$D^+u(x_0) = \{p \in \mathbb{R}^n;\ \limsup_{x \to x_0} [u(x) - u(x_0) - p \cdot (x - x_0)]|x - x_0|^{-1} \leqq 0\},$$

$$D^{-1}u(x_0) = \{p \in \mathbb{R}^n;\ \liminf_{x \to x_0} [u(x) - u(x_0) - p \cdot (x - x_0)]|x - x_0|^{-1} \geqq 0\}.$$

Then, $u \in C(G)$ is a viscosity solution if

$$|p| \leqq \rho \quad \text{for any } x_0 \in G, \quad p \in D^+u(x_0),$$

$$|p| \geqq \rho \quad \text{for any } x_0 \in G, \quad p \in D^-u(x_0).$$

From this definition it follows that if $u \in W^{1,\infty}(G)$ is a viscosity solution then it is also a weak solution; i.e., it satisfies the equation $|\nabla u| = \rho$ almost everywhere in G.

A function $u \in C(G)$ is said to be *semiconcave* if for any $\delta > 0$ there exist $C_\delta > 0$ such that $u(x) - C_\delta |x|^2$ is concave on every convex subset of

$$G_\delta \equiv \left\{ x \in G; |x| < \frac{1}{\delta}, \operatorname{dist}(x, \partial G) > \delta \right\}.$$

A function $u \in C(G)$ is said to be *semisuperharmonic* if for any $\delta > 0$ there exist $C_\delta > 0$ such that $\Delta u \leqq C_\delta$ in G_δ in the distribution sense.

Note that if u is semiconcave, then it is also semisuperharmonic.

THEOREM 5.1. *Under the assumptions* (5.1)-(5.5) *and* (5.11), u_ε *is a viscosity solution to the eikonal equation* (5.16) *in* Q_ε. *Moreover,* u_ε *is semiconcave and semisuperharmonic in any component* $\tilde{Q}_\varepsilon$ *of* Q_ε, *and it is given by*

$$u_\varepsilon(x) = \inf_{u \in \partial \tilde{Q}_\varepsilon} \{\phi(y) + L(x, y)\} \quad \forall x \in \tilde{Q}_\varepsilon, \tag{5.17}$$

where

$$\begin{aligned} L(x, y) = \inf \{ \rho T; \exists \xi \in \operatorname{Lip}[0, T], \xi(0) = x, \\ \xi(T) = y, \xi(t) \in \tilde{Q}_\varepsilon \ \forall t \in (0, T), |\xi'(t)| \leqq 1 \text{ a.e.}\} \end{aligned} \tag{5.18}$$

and

$$\phi(y) = \begin{cases} u_0(y) & \textit{if } u \in \partial \tilde{Q}_\varepsilon \cap \partial Q, \\ b & \textit{if } y \in \partial \tilde{Q}_\varepsilon \backslash \partial Q. \end{cases} \tag{5.19}$$

Proof. In view of Proposition 5.2 in Lions [10, p. 137] if suffices to show that u_ε is the maximum element of the set

$$S_\varepsilon = \{ v \in W^{1,\infty}(\tilde{Q}_\varepsilon), |\nabla v| \leqq \rho \text{ in } \tilde{Q}_\varepsilon, \ v \leqq u_\varepsilon \text{ on } \partial \tilde{Q}_\varepsilon \}.$$

Now, using (5.13) we have, for any $v \in S_\varepsilon$,

$$-\int_{\tilde{Q}_\varepsilon} \operatorname{div} \theta_\varepsilon (-u_\varepsilon + v)^+ \, dx = \int_{\tilde{Q}_\varepsilon} \theta_\varepsilon \nabla (-u_\varepsilon + v)^+ \, dx = \int_{\tilde{Q}_\varepsilon \cap \{-u_\varepsilon + v > 0\}} \theta_\varepsilon (-\nabla u_\varepsilon + \nabla v) \leqq 0.$$

Since, by (5.12), (5.14), $-\operatorname{div} \theta_\varepsilon \geqq c > 0$ on Q_ε, it follows that $(-u_\varepsilon + v)^+ = 0$. Thus $u_\varepsilon \geqq v$, which means that

$$u_\varepsilon(x) = \sup_{v \in S_\varepsilon} v(x) \quad \forall x \in \tilde{Q}_\varepsilon;$$

i.e., u_ε is the maximum element of S_ε.

Remark 5.1. By the proof of Theorem 1.4 in [3] and by Lemma 2.2, $u^* = \lim_{\varepsilon \to 0} u_\varepsilon$ is a viscosity solution of

$$|\nabla u| = \rho \quad \text{in } Q_0 = \{ x \in Q, u^*(x) < b \}.$$

If we knew that the boundaries $\partial \tilde{Q}_\varepsilon \backslash \partial Q$ are smooth (cf. [10, Remark 2.5, p. 70]) uniformly in ε, then we can apply Theorem 2.3 of [10, p. 66] to deduce $\Delta u_\varepsilon \leqq C'$ in compact subsets Q' of Q_0 with C' depending on Q' by not on ε (ε small enough). It follows that u^* is semisuperharmonic in every component $\tilde{Q}_0$ of Q_0. Hence, by uniqueness of semisuperharmonic viscosity solutions [10, p. 82] we conclude that u^* is the unique semisuperharmonic viscosity solution to (5.16) with Dirichlet data $u = u_0$ on $\partial Q \cap \partial \tilde{Q}_0$, $u = b$ on $\partial \tilde{Q}_0 \backslash \partial Q$; furthermore, the representation (5.17)-(5.19) extends to u^*. Note that this does not imply uniqueness of u^*, because Q_0 depends on u^*.

We will now use Theorem 5.1 to compute u^* in some simple cases.

Observe first that any component $\tilde{Q}_\varepsilon$ of Q_ε must satisfy

$$\partial \tilde{Q}_\varepsilon \cap \partial Q \neq \emptyset.$$

Indeed, otherwise we get from (5.17)-(5.19)

$$b > u_\varepsilon(x) = b + \inf_{y \in \tilde{Q}_\varepsilon} L(x, y) \quad \forall x \in \tilde{Q}_\varepsilon,$$

which is a contradiction. It follows that

(5.20) if ∂Q is connected, then Q_ε is connected.

Let $d(x) = \operatorname{dist}(x, \partial Q)$ for $x \in Q$. Set also

$$Q^* = \{x \in Q; \text{ there exists a unique point } y \in \partial Q \text{ such that } d(x) = |x - y|\},$$

$$d_0 = \min\{d(x); x \in \partial Q^* \backslash \partial Q\}, \qquad Q^*_\delta = \{x \in Q, d(x) < \delta\}.$$

We will assume for simplicity that ∂Q is connected and $u_0(x) = \text{const.} = u_0$. Since $u_\varepsilon(x) < b$ in Q_ε the "inf" in (5.17) cannot be attained for $y \in \partial \tilde{Q}_\varepsilon \backslash \partial Q$. Hence, recalling also (5.20),

$$u_\varepsilon(x) = u_0 + \inf_{y \in \partial Q} L(x, y) \quad \forall x \in Q_\varepsilon. \tag{5.21}$$

For small δ, $Q_\varepsilon \supset Q^*_\delta$ and, therefore,

$$\inf_{y \in \partial Q} L(x, y) = \rho d(x); \tag{5.22}$$

consequently,

$$u_\varepsilon(x) = u_0 + \rho d(x), \qquad d(x) < \delta. \tag{5.23}$$

As long as $u_0 + \rho\delta < b$ and $\delta < d_0$ we can continue to increase δ in small steps while proving that $Q_\varepsilon \supset Q^*_\delta$ and that (5.22), (5.23) hold. Consequently,

$$u_\varepsilon(x) = u_0 + \rho d(x) \quad \text{in } x \in Q^*_{\delta_0} \quad \text{where } \delta_0 = \min\left\{d_0, \frac{b - u_0}{\rho}\right\}.$$

Letting $\varepsilon \to 0$ we obtain Corollary 5.2.

COROLLARY 5.2. *Let* (5.1)-(5.5) *hold. Thus any optimal control* u^* *satisfies*

$$u^*(x) = u_0 + \rho d(x) \quad \textit{in } Q^*_\delta, \qquad \delta_0 = \min\left\{d_0, \frac{b - u_0}{\rho}\right\}. \tag{5.24}$$

If, in particular, $(b - u_0/)\rho \leqq d_0$ *then* $u^*(x) = b$ *in* $Q \backslash Q^*_{\delta_0}$ *and thus* u^* *is uniquely determined.*

In the case where $Q = \{x; |x| < R\}$, Corollary 5.2 implies that u^* is uniquely determined by

$$u^*(x) = \begin{cases} u_0 + \rho(R - |x|) & \text{if } |x| > R - \dfrac{b - u_0}{\rho}, \\ b & \text{if } |x| \leqq R - \dfrac{b - u_0}{\rho}. \end{cases} \tag{5.25}$$

To apply this theorem to the electrochemical machining problem (see Remark 5.2 below), we choose a function ψ satisfying

$$\psi \in C^2(\bar{Q}), \qquad \frac{\partial \psi}{\partial \nu} = 0 \quad \text{on } \partial Q, \tag{5.26}$$

$$b < \psi_0 \leqq \psi(x) \leqq H \quad \text{if } x \in \bar{Q} \quad (\psi_0 = \inf \psi),$$

and introduce the surface

$$\Gamma=\{y=\psi(x),\ x\in Q\}.$$

We take

$$z^0(x,y)=R(y-\psi(x)). \tag{5.27}$$

Then $\partial z^0/\partial\nu=0$ on $\partial Q\times(0,H)$, (5.2) reduces to

$$R''(y-\psi(x))[1+|\nabla\psi(x)|^2-R'(y-\psi(x))\Delta\psi(x)\leqq f, \tag{5.28}$$

and (5.3), (5.4) reduce to

$$R(y-\psi(x))>1 \quad \text{in } \bar{Q}_H\cap\{a\leqq y\leqq b\}, \tag{5.29}$$

$$R(y-\psi(x))>0 \quad \text{in } \bar{Q}_H. \tag{5.30}$$

The last two inequalities are satisfied if

$$R(b-\psi_0)>1, \qquad R(H)>0, \tag{5.31}$$

Remark 5.2. In the electrochemical machining problem we want to choose the control so that the free boundary $y=\mu(x)$ will be as close as possible to $y=\psi(x)$. This is hard to do. Suppose instead that we can choose in § 5

$$z^0=\begin{cases}0 & \text{if } y>\psi(x),\\ A & \text{if } y<\psi(x) \quad (A>0).\end{cases} \tag{5.32}$$

The solution z of the variational inequality satisfies

$$z=0 \quad \text{if } y>\mu(x),$$
$$z>0 \quad \text{if } y<\mu(x)$$

and thus the statement "$\|z-z^0\|$ is as small as possible" is a weak version of the statement "$\|\mu-\psi\|$ is as small as possible" (with suitable norms). This is the motivation for working with the functional (5.1). Next, for any small $\delta>0$ and $\varepsilon>0$, take a C^2 monotone function $R(s)$ satisfying

$$R(s)=\begin{cases}\varepsilon & \text{if } s>\delta,\\ A & \text{if } s<0 \quad (A>1).\end{cases} \tag{5.33}$$

If we choose $f\equiv C^*$ with C^* sufficiently large, then (5.28) and (5.31) are satisfied. Note that this choice of R yields a function $z^0(x,y)=R(y-\psi(x))$ which is very close (say, in L^1-norm) to the desired function in (5.32).

Remark 5.3. It is clear that the previous discussion and results remain valid if we take

$$F(X,z)=|z-z^0(X)|^p \tag{5.34}$$

where $p\geqq 2$ and z^0 satisfies (5.2)-(5.5). With F defined by (5.34) the cost functional (2.2) is a more accurate estimate of the "distance" between the coincidence set of z^* and the set $\{y\geqq\psi(x)\}$.

6. Generalizations. The results of §§ 2-5 extend to other control sets and functionals. We briefly mention one example whereby the control set is

$$U=\{u\in W^{1,p}(Q),\ a\leqq u(x)\leqq b\}, \qquad p>n, \tag{6.1}$$

and

$$J(u)=\int_{\Omega_u}F(X,z_u)\,dX+\lambda\int_Q|\nabla u|^p\,dx \qquad (\lambda>0); \tag{6.2}$$

since $p > n$, by Sobolev's inequality it follows that the control functions belong to $C^{\alpha}(\bar{Q})$ where $\alpha = (1/n) - (1/p)$. Here, in the ε-problem we replace $\partial h_U(u_\varepsilon)$ in (3.15) by

$$\partial h_1(u_\varepsilon) - \lambda p \Delta_p u_\varepsilon,$$

where Δ_p is the p-Laplacian

$$\Delta_p u = \operatorname{div}\,(|\nabla u|^{p-1} \nabla u)$$

taken in the distribution sense, and $\partial h_1(u_\varepsilon)$ is characterized by (4.4), (4.3). Thus (5.10) yields, as $\varepsilon \to 0$,

$$(6.3) \qquad -\Delta_p u^* \geqq c - \eta \quad c \text{ a positive constant and } \eta \text{ as in (4.3).}$$

We can therefore state the following result for Q in R^n, $n \geqq 1$.

THEOREM 6.1. *If* (5.1)-(5.5) *hold, then for any minimizer* u^* *of* (6.2), *in the class* (6.1), *the inequality* (6.3) *holds*; *in particular,*

$$(6.4) \qquad -\Delta_p u^* \geqq c > 0 \quad \textit{in the open set } \{u < b\}.$$

For $n = 1$, $p = 2$ we get that $u^*(x)$ is a convex function.

Remark 6.1. We recall (using (2.10)) that the free boundary for (1.1) is smooth (see [5, pp. 177-179]); if $f \in C^{m+\alpha}$ then the free boundary is given by $u = \varphi(x)$ with $\varphi \in C^{m+1+\alpha}$.

Remark 6.2. The method of this paper can be used, in principle, for other free boundary problems (see [6], [7]). However, in some cases it is difficult to derive meaningful properties of the optimal control from the optimality conditions for the ε-problem. We describe one such case which arises in contact problems for elastic bodies. In a rectangle

$$R = \{0 < x < 1, \, 0 < y < H\}$$

the control variable is a curve

$$\Gamma_u = \{y = u(x), \, 0 < x < 1\},$$

where u belongs to the control set

$$U = \{u \in \operatorname{Lip}[0, 1], |u'(x)| \leqq \rho, \, u(0) = u_0, \, u(1) = u_1, \, b \leqq u(x) \leqq H\} \qquad (0 < b < H)$$

and z is the solution of the Signorini problem (see, for instance, [5], [9]):

$$-\Delta z = f \quad \text{in } \Omega_u,$$

$$z \geqq 0, \quad \frac{\partial z}{\partial \nu} \geqq 0, \quad z \frac{\partial z}{\partial \nu} = 0 \quad \text{on } \Gamma_u,$$

$$z = 0 \quad \text{on } \partial\Omega_u \backslash \Gamma_u,$$

where $\Omega_u = R \cap \{y < u(x)\}$. The functional to be minimized is

$$J(u) = \int_{\Omega_u} (z_u - z^0(x, y))^2 \, dx \, dy,$$

where z^0 is a given function. This problem was studied by Hlaváček and Nečas [8], who proved the existence of an optimal control u^*. Using our method we define the

ε-problem whereby analogously to (2.4), (2.5) we take

$$-\Delta z = f \quad \text{in } \Omega_{\mu_\varepsilon(u)},$$

$$\frac{\partial z}{\partial \nu} + \beta_\varepsilon(z) = 0 \quad \text{on } \Gamma_{\mu_\varepsilon(u)},$$

$$z = 0 \quad \text{on } \partial\Omega_{\mu_\varepsilon(u)} \backslash \Gamma_{\mu_\varepsilon(u)}.$$

Proceeding as before, we arrive at the inequality

$$\begin{aligned} &\theta'_\varepsilon - \left[p \left(\frac{z_x}{(1+(u')^2)^{1/2}} + \frac{u'}{1+(u')^2} \frac{\partial z}{\partial \nu} \right) \right] \\ (6.5) \qquad &\geqq -p \frac{u' z_{xy} - z_{yy}}{(1+(u')^2)^{1/2}} + p\beta'_\varepsilon z_y + (u - u^*) + (z - z^0)^2 + \eta_\varepsilon \quad \text{on } \Gamma_{\mu_\varepsilon(u)} \cap [u_\varepsilon > b], \end{aligned}$$

where $u = u_\varepsilon$, $z = z_\varepsilon$, $\partial h_U(u_\varepsilon) - \theta'_\varepsilon + \eta_\varepsilon$, and where p_ε is the solution of

$$-\Delta p = z^0 - z_\varepsilon \quad \text{in } \Omega_{\mu_\varepsilon(u)},$$

$$\frac{\partial p_\varepsilon}{\partial \nu} + \beta'_\varepsilon(z_\varepsilon) p_\varepsilon = 0 \quad \text{on } \Gamma_{u_\varepsilon},$$

$$p_\varepsilon = 0 \quad \text{on } \partial\Omega_{\mu_\varepsilon(u)} \backslash \Gamma_{u_\varepsilon}.$$

Here Γ_{u_ε} is given by $\{y = \mu_\varepsilon(u_\varepsilon)(x)\}$.

Remark 6.3. For optimal design in elliptic variational inequalities, from the point of view of sensitivity analysis and numerical methods, see [13] and the references given there.

REFERENCES

[1] V. BARBU, *Nonlinear Semigroups and Differential Equations Nonlinear in Banach Space*, Nordhoff, Leyden, the Netherlands, 1976.

[2] ———, *Control of Variational Inequalities*, Pitman, London, 1984.

[3] 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.

[4] C. M. ELLIOTT AND J. R. OCKENDON, *Weak and Variational Methods for Moving Boundary Problems*, Pitman, London, 1982.

[5] A. FRIEDMAN, *Variational Principles and Free Boundary Problems*, John Wiley, New York, 1982.

[6] ———, *Optimal control for parabolic variational inequalities*, SIAM J. Control Optim., 25 (1987), pp. 482-497.

[7] A. FRIEDMAN, S. HUANG, AND J. YONG, *Bang-bang control for the dam problem*, Appl. Math. Optim., 15 (1987), pp. 68-85.

[8] I. HLÁČEK AND J. NEČAS, *Optimization of the domain in elliptic unilateral boundary value problems by finite element method*, RAIRO Anal. Numer., 16 (1982), pp. 351-373.

[9] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, New York, 1980.

[10] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi Equations*, Pitman, London, 1982.

[11] J. A. MCGEOUGH AND J. A. RASMUSSEN, *On the derivation of the quasi-stead model in electrochemical machinery*, J. Inst. Math. Appl., 13 (1974), pp. 13-21.

[12] J. NEČAS, *Directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[13] P. NEITTAANMÄKI, J. SOKOLOWSKI, AND J. P. ZOLESIO, *Optimization of the domain in elliptic variational inequalities*, Appl. Math. Optim., 18 (1988), pp. 85-98.

[14] R. T. ROCKAFELLAR, *On the maximality of sums of nonlinear operators*, Trans. Amer. Math. Soc., 149 (1970), pp. 55-88.

[15] ———, *Integrals which are convex functions*, II, Pacific J. Math., 39 (1971), pp. 439-469.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 3, pp. 638-655, May 1991

OPTIMAL CONSUMPTION BY A BOND INVESTOR: THE CASE OF RANDOM INTEREST RATE ADAPTED TO A POINT PROCESS*

PETER LAKNER† AND ERIC SLUD‡

Abstract. The problem of optimizing the expected total discounted utility of the rate of consumption by an agent with initial wealth x who, at each instant, can choose a rate of consumption and who invests all his unconsumed wealth in a bond (bank) is studied. The bond is assumed to have a randomly varying interest rate with known probabilistic behavior. A general martingale principle is formulated, according to which the optimal consumption rate is expressed in terms of a positive martingale, if one can be found, satisfying an almost sure integral condition. This martingale will be characterized in the case where the stochastic history of the interest rate is adapted to (i.e., expressible in terms of) an underlying counting process. The problem studied can be viewed as a special case of optimal consumption problems in "incomplete markets," a terminology introduced to the financial literature by Harrison and Pliska [*Stochastic Process. Appl.*, 11 (1981), pp. 215-260].

Key words. interest rate process, discounted utility, incomplete markets, counting process, martingale, minimal filtration, ordinary differential equation

AMS(MOS) subject classifications. 93E20, 49A60

Introduction. The present paper treats the problem of optimal consumption by an agent who invests his unconsumed wealth in a single bond (bank) with a stochastic interest rate, and who does not invest in the stock market.

Several authors discussed the optimization problem in the case where the agent invests in the stock market as well. The initial significant results were achieved by Merton in [11] and [12], and the more general, explicit solutions are presented in [2], [7]-[10] using the "martingale technique," instead of the method of dynamic programming. In [7]-[10], the financial market consists of several stocks and a bond. There is an agent who invests in the financial market and consumes on a finite time-horizon $[0, T]$, and defines his consumption preferences with a strictly concave, sufficiently smooth "utility function"

$$U:[0,\infty)\mapsto \mathcal{R}\cup\{-\infty\}.$$

The intuitive idea is that a rate of consumption $c(t)$ at time t gives our agent $e^{-\beta t}U(c(t))$ amount of "utility," where $\beta>0$ is a discount factor, and the agent strives to maximize the expected total discounted utility of consumption

$$E\int_0^T e^{-\beta t}U(c(t))\,dt,$$

subject to the restriction that he is not to be in debt at the terminal time.

In the above-mentioned papers there is a crucial assumption which is referred to as the "completeness of the market." This is a simple mathematical condition imposed on the vector of stock price processes, which guarantees the uniqueness of an auxiliary probability measure under which the vector of stock price processes, discounted by the current bond price, is a martingale. It also follows from the "complete market"

* Received by the editors June 21, 1989; accepted for publication (in revised form) June 15, 1990. This research was supported by Office of Naval Research contract N00014-89-J-1051.

† Stern School of Business, Department of Statistics and Operations Research, New York University, New York, New York 10003.

‡ Mathematics Department, University of Maryland, College Park, Maryland 20742.

condition that under this auxiliary probability measure, every martingale can be represented as a stochastic integral with respect to the discounted stock price processes. Assuming that the market is complete, the optimization problem is solved explicitly in [7]-[10]. For more information about the notion of complete market, see [3] and [4], where it is defined and its use is illustrated in the theory of contingent claim pricing.

The question of contingent claim pricing and optimal consumption and investment remained open in the case where the market is *incomplete*. In the present paper we address the latter problem, assuming that our agent, whom we will call a consumer, does not invest in the stock market, only in the bond (bank). Since our market model has no stocks, and is therefore incomplete, our setting is a special case of the optimal consumption/investment problem in an incomplete market.

The "classical" approach of Merton to this problem (in [11] and [12]) would be to assume that the interest rate process is a diffusion, write down the corresponding Hamilton-Jacobi-Bellman equation of dynamic programming, and try to solve it for particular utility functions like $U(c)=\log c$, or $U(c)=c^{\delta}$ with a positive constant $0<\delta<1$. However, being unwilling to make the assumption that the interest rate is Markovian, we cannot follow this path. Instead, in § 2 we develop a general principle called the *martingale principle*, which is a sufficient condition for optimality. In the following part of the paper we present an application of this principle in the case where the underlying filtration, which models the flow of information, is generated by a counting process.

We can put our problem in a different framework, called the Hotelling problem [6], concerning the optimal rate of depletion of a resource which reproduces itself at a random exponential rate. Everything in the present paper can be applied to the Hotelling problem in continuous time by a change of terminology.

1. The optimization problem of a consumer. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $[0, T]$ a bounded time-horizon, with positive terminal time $T<\infty$. The flow of information is represented by a filtration $\{\mathcal{F}_t: 0\leqq t\leqq T\}$ in $\mathcal{F}$. The σ-algebra $\mathcal{F}_t$ is generated by all the information available as of time t, including additional random data observed at time zero. The *interest rate* $\{r(t), \mathcal{F}_t: 0\leqq t\leqq T\}$ is assumed to be an adapted, measurable process, bounded above by a fixed positive constant K_0 and bounded below by -1.

We denote by $p(t)$ the price of the bond at time t, relative to its price at time zero. It evolves according to the equation

$$dp(t)=r(t)p(t)\,dt, \qquad p(0)=1. \tag{1.1}$$

The solution process

$$p(t)=\exp\left\{\int_0^t r(s)\,ds\right\}, \qquad 0\leqq t\leqq T \tag{1.2}$$

is positive, adapted to $\mathcal{F}_t$, bounded above, and bounded away from zero. Indeed, by the boundedness conditions imposed on the interest rate, the *discount process*

$$\gamma(t)\equiv\exp\left\{-\int_0^t r(s)\,ds\right\}=\frac{1}{p(t)} \tag{1.3}$$

satisfies

$$e^{-K_0T}\leqq\gamma(t)\leqq e^{T}, \qquad 0\leqq t\leqq T. \tag{1.4}$$

A *consumption process* $\{c(t), \mathcal{F}_t: 0 \leqq t \leqq T\}$ is a nonnegative, adapted process satisfying the integrability condition

$$\int_0^T c(s)\, ds < \infty \quad \text{a.s.} \tag{1.5}$$

The nonnegative, $\mathcal{F}_t$-measurable random variable $c(t)$ denotes the rate of consumption at time t. The condition that the consumption process $c(t)$ be adapted corresponds to the requirement that the consumption rate $c(t)$ depends only on the time t and on the available information at that time.

We will study the behavior of a consumer who has initial wealth $x > 0$ at time zero. Denote by $x(t)$ the value at time t of the consumer's wealth, which is assumed to be invested in the bond. If the consumption plan $c(t)$ is financed exclusively by the initial wealth and by the bank, then $x(t)$ satisfies

$$dx(t) = (r(t)x(t) - c(t))\, dt, \qquad x(0) = x, \tag{1.6}$$

where the factor $r(t)x(t)$ is the rate of gain or loss on the invested wealth. By (1.4) and (1.5), the unique solution of (1.6) is

$$x(t) = \frac{1}{\gamma(t)}\left(x - \int_0^t \gamma(s)c(s)\, ds\right). \tag{1.7}$$

A consumption process $\{c(t): 0 \leqq t \leqq T\}$ is said to be *admissible* if the corresponding terminal wealth is nonnegative, i.e., if

$$x(T) \geqq 0 \quad \text{a.s.} \tag{1.8}$$

By (1.7), a consumption process $\{c(t): 0 \leqq t \leqq T\}$ is admissible if and only if it satisfies the *budget inequality*

$$\int_0^T \gamma(t)c(t)\, dt \leqq x \quad \text{a.s.} \tag{1.9}$$

A *utility function* $U: (0, \infty) \mapsto \mathcal{R}$ will be a continuously differentiable, strictly increasing, strictly concave function, which satisfies the condition

$$\lim_{c \to \infty} U'(c) = 0. \tag{1.10}$$

It follows that the derivative function $U': (0, \infty) \mapsto (0, U'(0))$ is positive and strictly decreasing with $U'(0) \equiv U'(0+) \leqq \infty$ well defined, and therefore admits a strictly decreasing, continuous inverse $I: (0, U'(0)) \mapsto (0, \infty)$. If $U'(0) < \infty$, then we extend the inverse continuously on all of $(0, \infty)$ by defining $I(y) = 0$ for $y \geqq U'(0)$. For technical reasons (see Lemma A.1 in the Appendix), we also assume that for a sufficiently large positive constant $C < \infty$ to be fixed later (following condition (A3) in the Appendix) and for each $\delta > 0$, there is a constant $A = A(\delta)$ such that for all constants $1/C \leqq C_1$, $C_2 \leqq C$,

$$\left| U'\left(C_1 I\left(\frac{z}{C_2}\right)\right) - U'\left(C_1 I\left(\frac{z'}{C_2}\right)\right)\right| \leqq A|z - z'| \quad \text{for all } z, z' \geqq \delta, \tag{1.11}$$

and there is a constant A_1 not depending on δ for which

$$U'\left(C_1 I\left(\left(\frac{z}{C_2}\right)\right)\right) \leqq A_1 z \quad \text{for all } z > 0. \tag{1.12}$$

Extend the utility function $U(\cdot)$ to all of $[0, \infty)$ by

$$U(0) \equiv \lim_{c \to 0+} U(c) \geqq -\infty, \tag{1.13}$$

with the convention that the value of the function U at zero may be negatively infinite. In order to define his consumption preferences, our consumer chooses a utility function $U(\cdot)$ which satisfies the above conditions and a discount rate $\beta \in \mathcal{R}$, with the intuitive idea that rate of consumption $c(t)$ at time t gives him $e^{-\beta t}U(c(t))$ amount of "utility." In this paper we take the discount rate β to be a fixed constant. All of the results proved here extend easily to the case where β is replaced by a nonrandom function $\beta(\cdot)$ which is uniformly bounded away from 0 and ∞, and where all discount factors $e^{-\beta t}$ are replaced by $\exp(-\int_0^t \beta(s)\,ds)$.

The optimization problem of our consumer is to maximize the value of

$$E\int_0^T e^{-\beta t}U(c(t))\,dt \tag{1.14}$$

over all admissible consumption processes which satisfy the inequality

$$E\int_0^T e^{-\beta t}U^-(c(t))\,dt<\infty, \tag{1.15}$$

where $U^-(c)=-U(c)$ if $U(c)<0$, and $U^-(c)=0$ otherwise. Condition (1.15) guarantees that the expression in (1.14) is well defined. We denote by $\mathcal{A}(x)$ the class of admissible consumption policies that satisfy (1.15). The *value function* of this optimization problem is given by

$$V(x)\equiv\sup\left\{E\int_0^T e^{-\beta t}U(c(t))\,dt\colon \{c(t),0\leqq t\leqq T\}\in\mathcal{A}(x)\right\}. \tag{1.16}$$

Remark. The class $\mathcal{A}(x)$ is nonempty, since the constant consumption process

$$\bar{c}(t)=\bar{c}\equiv\frac{x}{T}e^{-T}$$

trivially satisfies (1.15), and (1.9) follows from (1.4).

For every admissible consumption process, (1.4) and (1.9) imply

$$\int_0^T c(t)\,dt\leqq x\,e^{K_0T},$$

and by Jensen's inequality

$$\int_0^T e^{-\beta t}U(c(t))\,dt\leqq\int_0^T U(c(t))\,dt\leqq TU\left(\int_0^T\frac{1}{T}c(t)\,dt\right)\leqq TU\left(\frac{x}{T}e^{K_0T}\right).$$

It follows that the value function of our optimization problem is finite: $V(x)<\infty$. Any consumption process $c(t)$ for which the expected discounted utility (1.14) achieves the value $V(x)$ is called *optimal.*

The following lemma shows that the optimal consumption process is unique.

LEMMA 1.1. *Let $c_1(t)$, $c_2(t)$ be two consumption processes in $\mathcal{A}(x)$, and assume that both are optimal. Then the Lebesgue $\times$ P measure of the set $\{(t,\omega)\in[0,T]\times\Omega\colon c_1(t,\omega)\neq c_2(t,\omega)\}$ is zero.*

Proof. Define

$$c_3(t)\equiv\tfrac{1}{2}(c_1(t)+c_2(t)).$$

By the convexity of the function $U^-(\cdot)$, the consumption process $c_3(t)$ satisfies (1.15). Obviously $c_3(t)$ satisfies (1.9), so it belongs to the class $\mathcal{A}(x)$. By the strict concavity of $U(\cdot)$ and the optimality of $c_1(t)$ and $c_2(t)$,

$$E\int_0^T e^{-\beta t}U(c_3(t))\,dt \geqq \frac{1}{2}\left(E\int_0^T e^{-\beta t}U(c_1(t))\,dt + E\int_0^T e^{-\beta t}U(c_2(t))\,dt\right) = V(x).$$

Since the last inequality cannot be strict, the lemma follows. □

2. A martingale principle. This section concerns a general principle of optimality for the optimization problem posed in § 1. The next result could be extended in an obvious way to the case $T=\infty$.

THEOREM 2.1. *Suppose that there exists a positive* $\mathcal{F}_t$*-martingale* $\{Y(t)\colon 0\leqq t\leqq T\}$ *which satisfies the identity*

$$\int_0^T \gamma(t)I(e^{\beta t}\gamma(t)Y(t))\,dt = x \quad a.s. \tag{2.1}$$

Then the consumption process

$$c^*(t) = I(e^{\beta t}\gamma(t)Y(t)) \tag{2.2}$$

belongs to the class $\mathcal{A}(x)$, *and is optimal for the consumer.*

Proof. The process $c^*(t)$ satisfies (1.9) trivially, since

$$E\int_0^T \gamma(t)c^*(t)\,dt = x \quad a.s. \tag{2.3}$$

Let $\{c(t)\colon 0\leqq t\leqq T\}$ be an arbitrary consumption process in the class $\mathcal{A}(x)$; by the remark following (1.16), such a process exists. Now we must show that $c^*(t)$ satisfies (1.15), and that

$$E\int_0^T e^{-\beta t}U(c^*(t))\,dt \geqq E\int_0^T e^{-\beta t}U(c(t))\,dt. \tag{2.4}$$

It follows from the concavity of $U(\cdot)$ that for all $c\geqq 0$ and $y>0$,

$$U(I(y)) \geqq U(c) + y(I(y)-c). \tag{2.5}$$

Substitute $y = e^{\beta t}\gamma(t)Y(t)$ and $c=c(t)$ in (2.5) to see that

$$U(c^*(t)) \geqq U(c(t)) + e^{\beta t}\gamma(t)Y(t)(c^*(t)-c(t)),$$

or, equivalently,

$$e^{-\beta t}U(c^*(t)) \geqq e^{-\beta t}U(c(t)) + \gamma(t)Y(t)(c^*(t)-c(t)). \tag{2.6}$$

In order to show that $c^*(t)$ satisfies (1.15) and (2.4), it suffices to show that

$$E\int_0^T \gamma(t)Y(t)c(t)\,dt \leqq E\int_0^T \gamma(t)Y(t)c^*(t)\,dt < \infty. \tag{2.7}$$

We define the new probability measure $\tilde{P}$ on $\mathcal{F}_T$ by

$$\tilde{P}(A) = \frac{1}{E[Y(0)]}E[1(A)Y(t)], \qquad A\in\mathcal{F}_t, \quad t\in(0,T]. \tag{2.8}$$

Since $Y(t)$ is a martingale, the above definition is consistent, and $\tilde{P}$ is a bona fide probability measure on $\mathcal{F}_T$, equivalent with P. We denote by $\tilde{E}$ the corresponding expectation operator. Using (2.3) and (1.9) we can write

$$\begin{aligned}E\left[\int_0^T \gamma(t)Y(t)c(t)\,dt\right] &= E[Y(0)]\tilde{E}\left[\int_0^T \gamma(t)c(t)\,dt\right] \leqq E[Y(0)]x\\ &= E[Y(0)]\tilde{E}\left[\int_0^T \gamma(t)c^*(t)\,dt\right]\\ &= E\left[\int_0^T \gamma(t)Y(t)c^*(t)\,dt\right] < \infty,\end{aligned}$$

which completes the proof of the theorem. □

The optimization problem has been reduced to finding a positive $\mathcal{F}_t$-martingale $Y(t)$ which satisfies (2.1).

3. Examples with history generated by a point process. We formulate the interest rate process in terms of a simple (nonexplosive, right-continuous, unit-jump) counting process $N(t)$, which counts the cumulative number up to time t of discrete events (such as changes in commercial-bank prime rate or federal reserve discount rate) related to major interest-rate changes. Let us generate the initial σ-algebra $\mathcal{F}_0$, which is generally not trivial, by information observed as of time zero, together with all negligible events in the probability space $(\Omega, \mathcal{F}, P)$. Then the interest rate process $r(t)$ and the corresponding discount process $\gamma(t)$ are adapted to

$$\mathcal{F}_t \equiv \mathcal{F}_0 \cup \mathcal{F}_t^N = \sigma(\mathcal{F}_0, \{N(u): 0 \leqq u \leqq t\}),$$

which is a "minimal right-continuous filtration" for the counting process $N(\cdot)$, and satisfies the "usual conditions" [1, Thm. T25, p. 304, and Thm. T35, p. 309].

Our primary assumptions are that $N(T)$ is bounded, i.e., that there exists a known constant n such that

$$N(T) \leqq n \tag{3.1}$$

and that $N(t)$ admits a predictable $\mathcal{F}_t$ intensity $\lambda(t)$ such that for some constants $0 < b < B < \infty$ and all $t \in [0, T]$,

$$b1_{\{N(t)<n\}} \leqq \lambda(t)1_{\{N(t)<n\}} \leqq B1_{\{N(t)<n\}}. \tag{3.2}$$

Let us assume first that we have a positive, $\mathcal{F}_t$-martingale $Y(t)$ satisfying (2.1), in order to derive necessary conditions. Then, by Theorem T20 of [1, p. 302], this martingale has a right-continuous modification, also denoted $Y(t)$, which is again a positive $\mathcal{F}_t$-martingale satisfying (2.1). The fundamental representation theorem for right-continuous $\mathcal{F}_t$-martingales [1, Thm. T9, p. 64] guarantees the existence of a predictable process $f(t)$ satisfying

$$\int_0^T |f(s)|\lambda(s)\,ds < \infty \quad \text{a.s.,} \tag{3.3}$$

such that

$$Y(t) = Y(0) + \int_0^t f(s)(dN(s) - \lambda(s)\,ds), \qquad 0 \leqq t \leqq T. \tag{3.4}$$

The integral on the right-hand side of (3.4) is a Riemann–Stieltjes integral.

Let $\tau_0=0$, and $\tau_1, \cdots, \tau_n$ denote the successive jump times of $N(t)$. Then every $\mathscr{F}_t$ adapted right-continuous process $\xi(t)$ can be represented almost surely on each event $\{\omega: \tau_k(\omega) \leqq t < \tau_{k+1}(\omega)\}$ for $k=0, \cdots, n-1$ as $\xi_k(t \mid \tau_1, \cdots, \tau_k)$ for some $\mathscr{F}_0$ measurable random element $\xi_k(\cdot \mid \cdot, \cdots, \cdot)$ of the space of measurable functions of $k+1$ arguments which are right-continuous in the first argument. For the rest of this paper, we adopt the notational convention that the ($\mathscr{F}_t$ predictable) processes $\gamma(t)$, $\lambda(t)$, $f(t)$, etc., are represented on random intervals $\tau_k(\omega) \leqq t < \tau_{k+1}(\omega)$ by functions denoted with the same letter written with subscripts and arguments ($\gamma_k(t \mid \tau_1, \cdots, \tau_k)$, $\lambda_k(t \mid \tau_1, \cdots, \tau_k)$, etc.).

3.1. The one-jump case. In this section, assume that $N(t)$ has at most one jump up to time T, i.e.,

$$N(T) \leqq 1 \quad \text{a.s.} \tag{3.5}$$

Then we have the following representations:

$$\lambda(t)=\lambda_0(t) 1_{\{t \leqq \tau_1\}}, \tag{3.6}$$

$$\gamma(t)=\gamma_0(t) 1_{\{t \leqq \tau_1\}}+\gamma_1(t \mid \tau_1) 1_{\{t>\tau_1\}}, \tag{3.7}$$

$$f(t)=f_0(t) 1_{\{t \leqq \tau_1\}}+f_1(t \mid \tau_1) 1_{\{t>\tau_1\}}. \tag{3.8}$$

It follows from (3.6)-(3.8) and (3.4) that on the event $\{t<\tau_1\}$

$$Y(t)=Y(0)-\int_0^t f_0(s) \lambda_0(s)\, ds, \tag{3.9}$$

whereas on the event $\{t \geqq \tau_1\}$

$$Y(t)=Y(0)+f_0(\tau_1)-\int_0^{\tau_1} f_0(s) \lambda_0(s)\, ds. \tag{3.10}$$

Since $f_1(\cdot)$ does not appear in (3.9) and (3.10), the only unknown function will be $f_0(\cdot)$. Consider the function

$$\alpha(t)=\int_0^t f_0(s) \lambda_0(s)\, ds, \qquad 0 \leqq t \leqq T. \tag{3.11}$$

By (2.1), almost surely on the event $\{\tau_1 \geqq T\}$

$$\int_0^T \gamma_0(t) I(e^{\beta t} \gamma_0(t)(Y(0)-\alpha(t)))\, dt=x, \tag{3.12}$$

whereas on $\{\tau_1<T\}$ almost surely

$$\begin{aligned} &\int_0^{\tau_1} \gamma_0(t) I(e^{\beta t} \gamma_0(t)(Y(0)-\alpha(t)))\, dt \\ &\quad+\int_{\tau_1}^T \gamma_1(t \mid \tau_1) I(e^{\beta t} \gamma_1(t \mid \tau_1)(Y(0)-\alpha(\tau_1)+f_0(\tau_1)))\, dt=x. \end{aligned} \tag{3.13}$$

By assumption (3.2), the support of τ_1 contains all of $[0, T]$, and the event $\{\tau_1 \geqq T\}$ has positive probability. Therefore, (3.12) and (3.13) must hold almost surely on $(\Omega, \mathscr{F}_0, P)$, the latter for Lebesgue-almost every $\tau_1 \in (0, T)$. Here, as throughout the rest of the paper, all equations involving functions $\lambda_k(t \mid \tau_1, \cdots, \tau_k)$, $\gamma_k(t \mid \tau_1, \cdots, \tau_k)$, etc., for $k \geqq 0$, should be understood as almost sure equalities among $\mathscr{F}_0$ measurable random variables or random elements of spaces of measurable functions (cf. the discussion immediately preceding § 3.1).

Define a function

$$Q(u\,|\,t,T)=\int_t^T \gamma_1(s\,|\,t)I(u\,e^{\beta s}\gamma_1(s\,|\,t))\,ds,\qquad 0<u<\infty,\quad 0\leqq t\leqq T. \tag{3.14}$$

It follows from the dominated convergence theorem, the monotone convergence theorem, and (1.4) that for each $t\in[0,T)$, the mapping $Q(\cdot\,|\,t,T)$ is continuous, strictly decreasing, and positive on an interval $(0,\Delta)$ for some constant $\Delta\equiv\Delta(t,T)\in(0,\infty)$; that $Q(u\,|\,t,T)=0$ if $\Delta<\infty$ and $u\geqq\Delta$; and that

$$\lim_{u\to\infty} Q(u\,|\,t,T)=0, \tag{3.15}$$

$$\lim_{u\to 0} Q(u\,|\,t,T)=\infty. \tag{3.16}$$

Therefore, $Q(\cdot\,|\,t,T)$ has a strictly decreasing, continuous pseudoinverse

$$\bar{Q}(v\,|\,t,T)=\inf\{u>0\colon Q(u\,|\,t,T)>v\}, \tag{3.17}$$

which satisfies

$$Q(\bar{Q}(v\,|\,t,T),t,T)=v,\qquad 0<v<\infty. \tag{3.18}$$

Rewrite (3.13) in the form

$$\int_0^t \gamma_0(s)I(e^{\beta s}\gamma_0(s)(Y(0)-\alpha(s)))\,ds+Q(Y(0)-\alpha(t)+f_0(t)\,|\,t,T)=x, \tag{3.19}$$

valid for almost every $t\in(0,T)$, and define for $t\in[0,T]$ the functions

$$h(t)=\exp\left\{-\int_0^t \lambda_0(s)\,ds\right\}, \tag{3.20}$$

$$Z_1(t)=h(t)(Y(0)-\alpha(t))>0, \tag{3.21}$$

and

$$Z_2(t)=x-\int_0^t \gamma_0(s)I\left(e^{\beta s}\gamma_0(s)\frac{Z_1(s)}{h(s)}\right)ds. \tag{3.22}$$

Then $h(\cdot)$ is differentiable for almost every $t\in(0,T)$, and

$$\dot{h}(t)=-\lambda_0(t)h(t)\quad\text{a.e. } t\in(0,T). \tag{3.23}$$

Now, by (3.21), (3.11), and (3.23), for almost every $t\in(0,T)$

$$\dot{Z}_1(t)=-\lambda_0(t)h(t)(Y(0)-\alpha(t)+f_0(t)), \tag{3.24}$$

and by (3.19), (3.21), (3.22), (3.24), for almost every $t\in(0,T)$

$$Q\left(-\frac{\dot{Z}_1(t)}{\lambda_0(t)h(t)}\,|\,t,T\right)=Z_2(t). \tag{3.25}$$

For Lebesgue almost every $t\in(0,T)$, (3.22) yields

$$\dot{Z}_2(t)=-\gamma_0(t)I\left(e^{\beta t}\gamma_0(t)\frac{Z_1(t)}{h(t)}\right) \tag{3.26}$$

and (3.21), (3.22), and (3.12) imply the boundary conditions

$$Z_1(0)=Y(0),\quad Z_2(0)=x,\quad Z_2(T)=0. \tag{3.27}$$

We summarize our results so far in the following theorem.

THEOREM 3.1. *Suppose that there exists a positive $\mathcal{F}_t$-martingale Y_t which satisfies* (2.1). *Then the functions $Z_1(\cdot)$, $Z_2(\cdot)$, given by* (3.21) *and* (3.22), *are absolutely continuous on* $[0, T]$ *and satisfy the differential equations* (3.25)-(3.26) *for almost every* $t \in (0, T)$, *with the boundary conditions* (3.27).

The converse of the preceding theorem is still more interesting for our purposes.

THEOREM 3.2. *Suppose that there exist two nonnegative, absolutely continuous functions $Z_1, Z_2 : [0, T] \mapsto [0, \infty)$, which for almost every $t \in (0, T)$ are positive and satisfy the* ODE *system*

$$(3.28)\qquad (\dot{Z}_1(t), \dot{Z}_2(t)) = \left(-\lambda_0(t)h(t)\bar{Q}(Z_2(t) \mid t, T), -\gamma_0(t)I\left(e^{\beta t}\frac{\gamma_0(t)}{h(t)} Z_1(t)\right)\right),$$

with the boundary conditions

$$(3.29)\qquad Z_1(0) < \infty, \quad Z_1(T) > 0, \quad Z_2(0) = x, \quad Z_2(T) = 0.$$

Then the process $\{Y(t) : 0 \leqq t \leqq T\}$ defined by

$$(3.30)\qquad Y(t) = \frac{Z_1(t)}{h(t)} \quad \text{on the event } \{0 < t < \tau_1\},$$

$$(3.31)\qquad Y(t) = \bar{Q}(Z_2(\tau_1) \mid \tau_1, T) \quad \text{on the event } \{\tau_1 \leqq t\} \cap \{\tau_1 < T\}$$

and defined arbitrarily on the null event $\{\tau_1 = T\}$, is a positive $\mathcal{F}_t$-martingale satisfying (2.1).

Proof. Since $Z_1(\cdot)$ and $Z_2(\cdot)$ must be strictly decreasing and positive on $(0, T)$, the process $Y(t)$ is well defined and positive. Define the function $f_0 : [0, T] \mapsto \mathcal{R}$ by $f_0(T) = 0$ and, for $0 \leqq t \leqq T$

$$(3.32)\qquad f_0(t) = -\frac{Z_1(t)}{h(t)} + \bar{Q}(Z_2(t) \mid t, T) = -\frac{1}{\lambda_0(t)}\frac{d}{dt}\left(\frac{Z_1(t)}{h(t)}\right)$$

and define the predictable process

$$(3.33)\qquad f(t) = f_0(t)1_{\{t \leqq \tau_1\}}, \qquad 0 \leqq t \leqq T.$$

Using (3.23) and the absolute continuity of the functions $Z_1(\cdot)$ and $h(\cdot)$, it is easy to show that (3.4) holds. Since $Z_1(\cdot)$ is strictly decreasing, and $\lambda_0(\cdot)$ and $1/h(\cdot)$ are bounded, it follows from (3.32) that

$$(3.34)\qquad E\int_0^T |f(t)|\lambda(t)\,dt \leqq e^{bT}\left\{Z_1(0)\int_0^{\min(\tau_1, T)} \lambda_0(t)\,dt + Z_1(0)\right\} < \infty$$

and $Y(t)$ is indeed a martingale [1, Thm. T8(β), p. 27]. By the absolute continuity of $Z_2(\cdot)$, together with (3.14) and (3.18), (2.1) follows immediately. □

The next step is to show that the ordinary differential equation system (3.28)-(3.29) does have a solution. Although we state the result here, its proof relies on the more general lemmas stated and proved in the Appendix.

THEOREM 3.3. *Under the assumptions of this section, the almost everywhere system* (3.28)-(3.29) *of ordinary differential equations in Theorem* 3.2 *has a unique solution, and this solution $\underline{Z}(t) = \binom{Z_1(t)}{Z_2(t)}$ is absolutely continuous as a function of t and satisfies the system for all $t \in [0, T]$.*

Proof. We verify the hypotheses of Lemma A.1 in the Appendix, where

$$F(z, t, T) \equiv -\dot{h}(t)\bar{Q}(z \mid t, T), \quad G_1(t) \equiv \gamma_0(t), \quad G_2(t) \equiv e^{\beta t}\frac{\gamma_0(t)}{h(t)}.$$

First, assumption (A1) on the function $I(\cdot)$ follows immediately from the assumptions on the utility function U made in the paragraph (1.10)-(1.13), and the uniform boundedness assumption (A3) on the functions $G_i(t)$ follows immediately for fixed $T<\infty$ from the bounds (1.4) on $\gamma(t)$ together with the definition of $h(t)$. We next check assumption (A2), that F is locally Lipschitz in $I^{-1}(z/(T-t))$. First, (3.14) says that

$$I^{-1}\left(\frac{Q(v\,|\,t,T)}{T-t}\right)-I^{-1}\left(\frac{Q(u\,|\,t,T)}{T-t}\right)=I^{-1}\left(\frac{1}{T-t}\int_t^T \gamma_1(s\,|\,t)I(v\,e^{\beta s}\gamma_1(s\,|\,t))\,ds\right)$$
$$-I^{-1}\left(\frac{1}{T-t}\int_t^T \gamma_1(s\,|\,t)I(u\,e^{\beta s}\gamma_1(s\,|\,t))\,ds\right).$$

By Lemma A.0 of Appendix A, with $G_1(s)=\gamma_1(s\,|\,t)$ and $G_2(s)=G_1(s)\,e^{\beta s}$, there exist s_1 and s_2 between t and T such that the last expression has the form

$$I^{-1}(\gamma_1(s_1\,|\,t)I(v\,e^{\beta s_2}\gamma_1(s_2\,|\,t)))-I^{-1}(\gamma_1(s_1\,|\,t)I(u\,e^{\beta s_2}\gamma_1(s_2\,|\,t))).$$

By (1.4) and then (A1.1) with $C\geqq\max(e^{K_0T}, e^{(\beta+1)T})$, we find that for all $u, v\geqq\delta^*\equiv I^{-1}(e^{(\beta+1)T}I(\delta\,e^{-K_0T}))$, the absolute value of the last displayed difference is at least $|u-v|/A(\delta)$. That is,

$$\left|I^{-1}\left(\frac{Q(v\,|\,t,T)}{T-t}\right)-I^{-1}\left(\frac{Q(u\,|\,t,T)}{T-t}\right)\right|\geqq\frac{|u-v|}{A(\delta)}.$$

Since $Q(u\,|\,t,T)\leqq e^{-K_0T}(T-t)I(\delta^*)$ implies that $u\geqq\delta^*$, we find that the inverse function $\bar{Q}(z\,|\,t,T)$ has the property

$$|\bar{Q}(z\,|\,t,T)-\bar{Q}(y\,|\,t,T)|\leqq A(\delta)\left|I^{-1}\left(\frac{z}{T-t}\right)-I^{-1}\left(\frac{y}{T-t}\right)\right|$$

for $y, z\leqq e^{-K_0T}(T-t)I(\delta^*)$. Similarly, it follows from (1.4) and (3.14) that $Q(u\,|\,t,T)\geqq e^{-K_0T}(T-t)I(u\,e^{(\beta+1)T})$, so that by (3.17)

$$\bar{Q}(z\,|\,t,T)\leqq e^{K_0T}I^{-1}\left(e^{-T}\frac{z}{T-t}\right)\quad\text{for all } z.$$

Thus, if we restrict T to lie in an interval $[0, T_*]$, then there exists a constant D $(=\exp(K_0T_*))$ such that for all positive z,

$$\bar{Q}(z\,|\,t,T)\leqq DI^{-1}\left(\frac{z}{(T-t)D}\right).$$

Since $|\dot{h}(t)|$ is uniformly bounded away from 0 and ∞, assumption (A2) of the Appendix holds for $F(z,t,T)$ as defined above. The conclusion of our Theorem is now part of Lemma A.1. □

3.2. The *n*-jump case. So far in this section, we have proved that a unique optimal consumption policy, determined by the "martingale principle" and an ordinary differential equation, exists under the very special assumption that the counting process $N(t)$ to which the interest-rate process is adapted has at most one jump in the bounded interval $[0, T]$. We now return to the case (3.1), where $N(T)$ is assumed bounded by the known integer n, and prove by a backward induction that the martingale $Y(t)$ determining an optimal policy is determined by recursively solving at most n ordinary differential equations.

Observe first that under (3.1), on the event $\{\omega\colon \tau_{n-1}(\omega) < T\}$, if the consumer has used the admissible policy $c(t)$ to guide his consumption on the time interval $[0, \tau_{n-1}]$, then his wealth as of time τ_{n-1} is $x_{n-1} \equiv x - \int_0^{\tau_{n-1}} c(t)\gamma(t)\,dt$. The restarted counting process $N(\tau_{n-1}+s) - N(\tau_{n-1})$ on the $\mathcal{F}_{\tau_{n-1}}$ measurable random time interval $[0, T-\tau_{n-1})$ has almost surely at most one jump. The problem of optimal consumption on the remaining interval $(\tau_{n-1}, T]$, posed conditionally given $\mathcal{F}_{\tau_{n-1}}$, is precisely the problem which we have just solved in the previous section, where initial wealth is replaced by x_{n-1}, the time horizon T by $T - \tau_{n-1}$, the time-variable t by $s = t - \tau_{n-1}$, the interest rate process by $\gamma(\tau_{n-1}+s)$, and the function $I(z)$ by $\tilde{I}(z) \equiv I(z\,e^{\beta\tau_{n-1}})$. In the present setting, Theorem 3.2 identifies the optimal consumption rate $c_{n-1}(t\,|\,\tau_1, \cdots, \tau_{n-1})$ for $\tau_{n-1} \leqq t \leqq T$, conditionally given $\mathcal{F}_{\tau_{n-1}}$, as follows.

Replace x by $x_{n-1} = x - \int_0^{\tau_{n-1}} c(t)\gamma(t)\,dt$ and T by $T-\tau_{n-1}$ in (A4), and define functions F and G_i $(i = 1, 2)$ in terms of the time-variable $s = t - \tau_{n-1}$ for $t \in [\tau_{n-1}, T]$ by

$$
\begin{aligned}
&F(z, t-\tau_{n-1}, T-\tau_{n-1})\\
&\quad = \lambda_{n-1}(t\,|\,\tau_1, \cdots, \tau_{n-1}) \exp\left(-\int_{\tau_{n-1}}^{t} \lambda_{n-1}(y\,|\,\tau_1, \cdots, \tau_{n-1})\,dy\right)\\
&\qquad \cdot \inf\left\{u > 0\colon \int_t^T \gamma_{n-1}(y\,|\,\tau_1, \cdots, \tau_{n-1}) I(u\,e^{\beta y}\gamma_{n-1}(y\,|\,\tau_1, \cdots, \tau_{n-1}))\,dy > z\right\},
\end{aligned}
$$

$$G_1(s) = \gamma_{n-1}(s+\tau_{n-1}\,|\,\tau_1, \cdots, \tau_{n-1}),$$

$$G_2(s) = G_1(s) \exp\left(\beta s + \int_{\tau_{n-1}}^{s+\tau_{n-1}} \lambda_{n-1}(y\,|\,\tau_1, \cdots, \tau_{n-1})\,dy\right).$$

The differential equation system (A4.0) defined in this way with initial and terminal conditions

$$Z_1^{(n)}(0\,|\,x_{n-1}) = x_{n-1} \quad \text{and} \quad Z_2^{(n)}((T-\tau_{n-1})-)|x_{n-1}) = 0$$

has a unique solution $\underline{Z}^{(n)}(s\,|\,x_{n-1})$ for $0 \leqq s \leqq T-\tau_{n-1}$, and $Z_1^{(n)}(s\,|\,x_{n-1})$ is monotonically decreasing in the parameter x_{n-1}. For $\tau_{n-1} \leqq t$, define the positive $\mathcal{F}_{\tau_{n-1}}$ measurable random function $Y_{n-1}(\cdot\,|\,\cdot, \cdots, \cdot) \equiv Y_{n-1}(\cdot\,|\,\cdot, \cdots, \cdot, x_{n-1})$ as in (3.30)-(3.31) by

$$
Y_{n-1}(t\,|\,\tau_1, \cdots, \tau_{n-1}, x_{n-1}) = \begin{cases} \exp\left(\int_0^t \lambda(v)\,dv\right) Z_1^{(n)}(t\,|\,x_{n-1}) & \text{if } t < \tau_n,\\[1ex] \exp\left(\int_0^{\tau_n} \lambda(v)\,dv\right) \dfrac{\dot{Z}_1^{(n)}(\tau_n\,|\,x_{n-1})}{\lambda(\tau_n)} & \text{if } t \geqq \tau_n. \end{cases}
$$

Then $Y_{n-1}(\tau_{n-1}+s\,|\,\tau_1, \cdots, \tau_{n-1}, x_{n-1})$ is a positive martingale in s, and the unique optimal consumption rate on $[\tau_{n-1}, T]$, after having used $c(\cdot)$ on $[0, \tau_{n-1})$, is

$$c_{n-1}(t\,|\,\tau_1, \cdots, \tau_{n-1}, x_{n-1}) = I(\gamma_{n-1}(t\,|\,\tau_1, \cdots, \tau_{n-1})\,e^{\beta t}\,Y_{n-1}(t\,|\,\tau_1, \cdots, \tau_{n-1}, x_{n-1})).$$

The only way in which the policy $c(\cdot)$ used on $[0, \tau_{n-1})$ enters into the expression for $c_{n-1}(t\,|\,\tau_1, \cdots, \tau_{n-1}, x_{n-1})$ is through the parameter $x_{n-1} = x - \int_0^{\tau_{n-1}} c(t)\gamma(t)\,dt$ denoting the consumer's wealth as of the random time τ_{n-1}. One of the immediate consequences of Lemma A.1 is that the function $Y_{n-1}(t\,|\,\tau_1, \cdots, \tau_{n-1}, x_{n-1})$ has a Lipschitz, nondecreasing dependence on the parameter $x_{n-1}/(T-\tau_{n-1})$ over bounded sets.

We next describe the backwards inductive step in solving for the martingale $Y(t)$ on $(\tau_{k-1}, \tau_k]$ after its trajectory on $(\tau_k, T]$ has been expressed as a nondecreasing locally Lipschitz function of $x_k/(T-\tau_k)$, where x_k denotes the consumer's wealth at random time τ_k.

THEOREM 3.4. *Suppose $1 \leqq k \leqq n$ is fixed and that the nonnegative $\mathscr{F}_t$ adapted random function $Y_k(t \mid \tau_1, \cdots, \tau_k, x_k)$ is nondecreasing and locally Lipschitz in $x_k/(T-\tau_k)$ on $(0, \infty)$, is uniformly bounded by $BI^{-1}(x_k/(T-\tau_k)B)$ for some constant B and, for $\tau_k \leqq t$, has the form*

$$\begin{aligned} Y_k(t \mid \tau_1, \cdots, \tau_k, x_k) \equiv {} & Y_k(\tau_k + \mid \tau_1, \cdots, \tau_k, x_k) \\ & - \int_{\tau_k}^{t \wedge \tau_{k+1}} f_k(y \mid \tau_1, \cdots, \tau_k, x_k) \lambda_k(y \mid \tau_1, \cdots, \tau_k)\, dy \\ & + 1_{\{\tau_{k+1} \leqq t\}} f_k(\tau_{k+1} \mid \tau_1, \cdots, \tau_k, x_k). \end{aligned} \tag{3.34$_k$}$$

Then there is a unique function $1_{\{\tau_{k-1} \leqq t\}} Y_{k-1}(t \mid \tau_1, \cdots, \tau_{k-1}, x_{k-1})$ satisfying the same properties and of the form (3.34_{k-1}), which is determined for $\tau_k = t \in (\tau_{k-1}, T]$ by

$$\begin{aligned} & Y_{k-1}(t \mid \tau_1, \cdots, \tau_{k-1}, x_{k-1}) \\ & \quad = Y_k\Bigg(t + \mid \tau_1, \cdots, \tau_{k-1}, t, x_{k-1} - \int_{\tau_{k-1}}^{t} \gamma_{k-1}(s \mid \tau_1, \cdots, \tau_{k-1}) \\ & \qquad \cdot I(e^{\beta s} \gamma_{k-1}(s \mid \tau_1, \cdots, \tau_{k-1}) Y_{k-1}(s \mid \tau_1, \cdots, \tau_{k-1}, x_{k-1}))\, ds \Bigg) \end{aligned} \tag{3.35}$$

and

$$\begin{aligned} x_{k-1} = \int_{\tau_{k-1}}^{T} & \gamma_{k-1}(s \mid \tau_1, \cdots, \tau_{k-1}) \\ & \cdot I(e^{\beta s} \gamma_{k-1}(s \mid \tau_1, \cdots, \tau_{k-1}) Y_{k-1}(s \mid \tau_1, \cdots, \tau_{k-1}, x_{k-1}))\, ds. \end{aligned} \tag{3.36}$$

Note. First, condition (3.34) says that the random functions $Y_j(t \mid \tau_1, \cdots, \tau_j, x_j)$ for $j = k$ and $k-1$ are nonrandom measurable functions of their arguments on the respective intervals $[\tau_k, \tau_{k+1})$ and $[\tau_{k-1}, \tau_k)$. In addition, (3.34) says that

$$\begin{aligned} & Y_j(\tau_{j+1} - \mid \tau_1, \cdots, \tau_j) - Y_j(\tau_{j+1} \mid \tau_1, \cdots, \tau_j) \\ & \quad = \frac{1}{\lambda(\tau_{j+1})} \frac{d}{dt} Y_j(t \mid \tau_1, \cdots, \tau_j) \big|_{t = \tau_{j+1}-}, \end{aligned} \tag{3.37}$$

where the displayed derivative makes sense for all $t < \tau_{j+1}$ and can be regarded as a left-derivative at τ_{j+1}. Next, since

$$\begin{aligned} x_k = x_{k-1} - \int_{\tau_{k-1}}^{t} & \gamma_{k-1}(s \mid \tau_1, \cdots, \tau_{k-1}) \\ & \cdot I(e^{\beta s} \gamma_{k-1}(s \mid \tau_1, \cdots, \tau_{k-1}) Y_{k-1}(s \mid \tau_1, \cdots, \tau_{k-1}, x_{k-1}))\, ds \end{aligned} \tag{3.38}$$

should denote the consumer's wealth as of time $\tau_k = t > \tau_{k-1}$, we recognize condition (3.35) simply as saying that the martingale $Y(\cdot)$ to be defined equal to $Y_{k-1}(\cdot \mid \tau_1, \cdots, \tau_{k-1}, x_{k-1})$ on $[\tau_{k-1}, \tau_k \wedge T]$ for $k = 1, 2, \cdots, n$, is right-continuous at each time τ_k which is less than T. Condition (3.36) recapitulates the requirement (2.1) for $Y(\cdot)$ on the event $[\tau_{k-1} < T \leqq \tau_k]$.

Proof. In order to characterize and solve for $Y_{k-1}(t \mid \tau_1, \cdots, \tau_{k-1})$ using (3.34)–(3.35), we will solve on $s = t - \tau_{k-1} \in [0, T - \tau_{k-1}]$ for $\underline{Z}^{(k)}(s)$ defined by

$$Z_1^{(k)}(t - \tau_{k-1}) = \exp\left(- \int_{\tau_{k-1}}^{t} \lambda_{k-1}(y \mid \tau_1, \cdots, \tau_{k-1})\, dy \right) Y_{k-1}(t \mid \tau_1, \cdots, \tau_{k-1}, x_{k-1}) 1_{\{\tau_k \geqq t\}}$$

$$\begin{aligned} & Z_2^{(k)}(t - \tau_{k-1}) \\ & \quad = x_{k-1} - \int_{\tau_{k-1}}^{t} I(e^{\beta y} \gamma_{k-1}(y \mid \tau_1, \cdots, \tau_{k-1}) Y_{k-1}(y)) \gamma_{k-1}(y \mid \tau_1, \cdots, \tau_{k-1})\, dy. \end{aligned}$$

According to (3.35)-(3.37) this vector function $\underset{\sim}{Z}^{(k)}(s)$ satisfies the almost-everywhere ordinary differential equation system (A4) in Lemma A.1 of Appendix A, where T in (A4) is replaced by $T-\tau_{k-1}$, x by x_{k-1}, and $I(z)$ by $\tilde{I}(z) \equiv I(z\, e^{\beta\tau_{k-1}})$, and where

$$\begin{aligned}F(z, s, T-\tau_{k-1}) = {} & \lambda_{k-1}(s+\tau_{k-1} \,|\, \tau_1, \cdots, \tau_{k-1}) \\ & \cdot \exp\left(-\int_{\tau_{k-1}}^{s+\tau_{k-1}} \lambda_{k-1}(y \,|\, \tau_1, \cdots, \tau_{k-1})\, dy\right) \\ & \cdot Y_k((s+\tau_{k-1})+ \,|\, \tau_1, \cdots, \tau_{k-1}, \tau_{k-1}+s, z), \\ G_1(s) = {} & \gamma_{k-1}(s+\tau_{k-1} | \tau_1, \cdots, \tau_{k-1}), \\ G_2(s) = {} & G_1(s) \exp\left(\beta s + \int_{\tau_{k-1}}^{\tau_{k-1}+s} \lambda_{k-1}(y \,|\, \tau_1, \cdots, \tau_{k-1})\, dy\right).\end{aligned}$$

The existence, uniqueness, and asserted regularity properties of the solution $\underset{\sim}{Z}^{(k)}(s)$ all follow immediately from Lemma A.1, the hypotheses of which hold by the assumptions of the present theorem. We note in conclusion that by the special form of (3.34), each of the random functions $Y_j(\tau_j+s \,|\, \tau_1, \cdots, \tau_j, x_j)$ is a martingale in $s \geqq 0$, and is a local martingale by [1, p. 27], with integrability verified exactly as in Theorem 3.2. □

It remains to tie together the construction of Y_{n-1} with the inductive step provided by Theorem 3.4. Note first that each of the systems of ordinary differential equations defining $\underset{\sim}{Z}^{(j)}(s)$ (and thereby defining Y_j) in the preceding proof involves only its own initial condition x_j and none of the other wealth-values x_k. Thus, after having obtained all the forms of $Y_j(t \,|\, \tau_1, \cdots, \tau_j, x_j)$ for $j = 0, \cdots, n-1$ as measurable functions of their arguments on $[\tau_j \leqq t < \tau_{j+1}]$, we proceed by *forward* induction as follows. First fix the total wealth $x_0 = x$, and define $Y(t) \equiv Y_0(t \,|\, x_0)$ for $0 \leqq t \leqq T$, $t < \tau_1$. Next, define inductively for $j = 1, 2, \cdots, n-1$, on the event $[\tau_j \leqq T]$,

$$\begin{aligned}x_j = x_{j-1} - \int_{\tau_{j-1}}^{\tau_j} & \gamma_{j-1}(s \,|\, \tau_1, \cdots, \tau_{j-1}) \\ & \cdot I(e^{\beta s}\gamma_{j-1}(s \,|\, \tau_1, \cdots, \tau_{j-1}) Y_{j-1}(s \,|\, \tau_1, \cdots, \tau_{j-1}, x_{j-1}))\, ds, \\ Y(t) = {} & Y_j(t \,|\, \tau_1, \cdots, \tau_j, x_j) \quad \text{for } \tau_j \leqq t \leqq T, \quad t < \tau_{j+1}.\end{aligned}$$

Use (3.37) to define $Y(t) = Y_{n-1}(\tau_n \,|\, \tau_1, \cdots, \tau_{n-1}, x_{n-1})$ for $\tau_n \leqq t \leqq T$ on the event $[\tau_n \leqq T]$. Then by construction and (3.37), $Y(\tau_j)$ agrees with $Y_{j-1}(\tau_j \,|\, \tau_1, \cdots, \tau_{j-1}, x_{j-1})$ for every τ_j which is less than or equal to T. The martingale property in s of each random function $Y_j(\tau_j+s \,|\, \tau_1, \cdots, \tau_j, x_j)$ implies immediately that the $\mathscr{F}_t$ adapted random function $Y(t)$ on $[0, T]$ is also a martingale (with respect to $\mathscr{F}_t$). Finally, the definition of x_j from x_{j-1}, together with the construction of the theorem, imply that $Y(t)$ satisfies (2.4) almost surely. The martingale principle of § 2 (Theorem 2.1) implies Theorem 3.5.

THEOREM 3.5. *Under assumptions* (3.1) *and* (3.2) *on the underlying counting process* $N(t)$, *for* $Y(t)$ *defined as in the present paragraph, the consumption rate* $c^*(t) \equiv I(e^{\beta t}\gamma(t) Y(t))$ *optimizes expected discounted utility* (1.13) *over all admissible consumption rates.*

4. Discussion and extensions. The main result of this paper, summarized in Theorem 3.5, says that under the condition (3.1) bounding a priori the number of jumps of $N(\cdot)$ on $[0, T]$, and under the regularity conditions assumed for the utility function, there exists a unique $\mathscr{F}_t$ adapted optimal consumption rate $c^*(t)$, which is constructed from the positive martingale $Y(t)$.

A number of possible interesting extensions of Theorem 3.5 are topics for further research. First, at least under some auxiliary assumption allowing one to prove a priori bounds on the optimal martingale $Y(t)$ or consumption rate $c^*(t)$, it seems likely that passage to limits as $n \to \infty$ should allow the bound n in (3.1) and (3.2) to be replaced by ∞. Second, it would be desirable to extend Theorem 3.5 to the case where the underlying process $N(t)$, to which both the interest rate and consumption rate processes are adapted, would be a multivariate or marked point process. If these first two extensions were accomplished, then the lower bound in (3.2) could be removed by replacing $N(t)$ by a multivariate counting process $N^*(t)$ with first component $N(t)$ and second component an independent Poisson process with intensity b. Other open problems connected with the martingale $Y(t)$ of this paper include: the relationship with the "value function," which plays such an important role in related treatments of interest rates adapted to diffusion processes, and the possibility of obtaining results for diffusion and more general interest-rate processes without jumps by passing to a high-intensity limit within the present framework.

Appendix A. Lemma on ordinary differential equations. The following summarizes the properties of equations (3.28)-(3.29) needed to show that each of these equations has in general a well-defined and unique solution. The results of this Appendix are self-contained, although for the applications in § 3, the function $I(\cdot)$ is the pseudo-inverse of the derivative $U'(\cdot)$ of the utility function $U(\cdot)$ of § 1.

(A1) For some positive ν which may be $+\infty$, $I:(0,\nu)\to(0,\infty)$ is a continuous and strictly decreasing function such that $I(0+)=\infty$ and $I(\nu)=0$. The function $I(\cdot)$ is defined $\equiv 0$ on $[\nu,\infty)$ if $\nu<\infty$. For some sufficiently large constant $C<\infty$ and each $\delta>0$, there are constants $A=A(\delta)$ such that for all constants $C^{-1}\leqq C_1, C_2 \leqq C$,

$$\left|I^{-1}\left(C_1 I\left(\frac{z}{C_2}\right)\right)-I^{-1}\left(C_1 I\left(\frac{z'}{C_2}\right)\right)\right|\leqq A|z-z'| \quad \text{for } z, z'\geqq\delta \tag{A1.1}$$

and there is a constant A_1 depending on C but not on δ for which

$$I^{-1}\left(C_1 I\left(\frac{z}{C_2}\right)\right)\leqq A_1 z \quad \text{for all } z>0. \tag{A1.2}$$

(A2) $F(z,t,T)$ is a nonnegative measurable function on $(0,\infty)\times\{(t,T): 0<t<T\leqq T_*\}$ for some $T_*<\infty$, which is nonincreasing in z with $F(0+,t,T)\leqq\infty$, and which is Lipschitz with respect to $I^{-1}(z/(T-t))$ on each set $\{(z,t,T): z/(T-t)\leqq\delta^{-1}\}$ for $\delta>0$. Moreover, there exists $D<\infty$ such that for all $z>0$ and $t<T$,

$$F(z,t,T)\leqq DI^{-1}\left(\frac{z}{(T-t)D}\right).$$

(A3) For $i=1,2$, $G_i(t)$ are positive continuous functions on $[0,T_*]$ such that for some $K_1<\infty$, $|\log G_i(t)|\leqq K_1$ for all $t\in[0,T_*]$.

In these three assumptions, T_* plays the role of a fixed upper bound on the time-horizon T, and the constant C in (A1.1) may and from now on will be fixed equal to $D\cdot\exp(K_1)$, with $D\geqq 1$.

LEMMA A.0. *For arbitrary $t<T$, there exist values s_1 and s_2 in $[t,T]$ such that*

$$\frac{1}{T-t}\int_t^T G_1(s)\binom{I(G_2(s)w)}{I(G_2(s)v)}\,ds = G_1(s_1)\binom{I(G_2(s_2)w)}{I(G_2(s_2)v)}.$$

Proof. Fix $t < T$, and let $v \leqq w$ be such that $I(G_2(s)v) > 0$ for some $s \in [t, T]$. Then the intermediate value theorem implies that there exists $s_2 \in [t, T]$ for which

$$\int_t^T G_1(s)\begin{pmatrix} I(G_2(s)w) \\ I(G_2(s)v) \end{pmatrix} ds \quad \text{is parallel to} \quad \begin{pmatrix} I(G_2(s_2)w)/I(G_2(s_2)v) \\ 1 \end{pmatrix},$$

i.e., for which

$$\frac{\int_t^T G_1(s) I(G_2(s)w)\, ds}{\int_t^T G_1(s)\{I(G_2(s)w) + I(G_2(s)v)\}\, ds} = \frac{I(G_2(s_2)w)}{I(G_2(s_2)w) + I(G_2(s_2)v)}.$$

Then the integral mean value theorem applied to

$$\frac{1}{T-t}\int_t^T G_1(s)\begin{pmatrix} I(G_2(s)w) \\ I(G_2(s)v) \end{pmatrix} \cdot \begin{pmatrix} I(G_2(s_2)w) \\ I(G_2(s_2)v) \end{pmatrix} ds$$

implies the existence of s_1. □

LEMMA A.1. *For each fixed* $T \leqq T_*$ *and* $x > 0$, *there exists a unique value* y_0 *such that the (unique) solution* $\underline{Z}(t) \equiv \binom{Z_1(t)}{Z_2(t)}$ *on* $(0, T)$ *in the almost-everywhere sense of the ordinary differential equation system*

$$\text{(A4.0)} \qquad \dot{\underline{Z}}(t) = \begin{pmatrix} -F(Z_2(t), t, t) \\ -G_1(t) I(G_2(t) Z_1(t)) \end{pmatrix}, \qquad \underline{Z}(0) = \begin{pmatrix} y_0 \\ x \end{pmatrix}$$

satisfies $Z_2(T-) = 0$, $Z_1(T-) > 0$. *Moreover,* $y_0 = y_0(T, x)$ *is continuous in both arguments, and is strictly decreasing and Lipschitz with respect to* x/T *on* $\{(t, x)\colon 0 \leqq T \leqq T_*, x/T \leqq \delta^{-1}\}$ *for each* $\delta > 0$. *In addition, there exists a finite positive constant* B *such that for all* z *and all* $T \leqq T_*$, $y_0(T, x) \leqq BI^{-1}(x/BT)$.

Proof. The main part of the proof consists in studying the existence and properties of the solution in almost-everywhere sense of

$$\text{(A4)} \qquad \dot{\underline{Z}}(t) = \begin{pmatrix} -F(Z_2(t), t, T) \\ -G_1(t) I(G_2(t) Z_1(t)) \end{pmatrix}, \qquad \underline{Z}(T-) = \begin{pmatrix} z \\ 0 \end{pmatrix}$$

as a function of $z > 0$. The strictly monotone dependence of each component of the right-hand side of (A4) will be used to establish the existence of a decreasing and absolutely continuous solution $Z_1(t)$. We omit the qualification "almost everywhere" in equalities and inequalities deduced from (A4).

The integrated form of (A4) tells that the function $W_1(t) \equiv Z_1(T-t) - z$ for $0 < t \leqq T$ must satisfy

$$\text{(A4}'\text{)} \quad W_1(t) = \int_{T-t}^T F\left(\int_s^T G_1(v) I(G_2(v)(W_1(T-v) + z))\, dv, s, T\right) ds, \qquad W_1(0) = 0.$$

Now fix $K = \exp(K_1) = k^{-1}$, and let $\mathscr{A}_T$ denote the set of nondecreasing functions $H(t)$ on $[0, T]$ such that $H(0) = 0$. Note that with our choice $C = D \exp(K_1)$, we have $C \geqq K$. For each $z > 0$, define the transformation R_z on $\mathscr{A}_T$ by

$$(R_z H)(t) \equiv \int_0^t F\left(\int_0^s G_1(T-v) I(G_2(T-v)(H(v) + z))\, dv, T-s, T\right) ds.$$

Then by (A1) and (A2), R_z is evidently monotone increasing in the sense that

$$H_1(t) \leqq H_2(t) \quad \text{for all } t \Rightarrow (R_z H_1)(t) \leqq (R_z H_2(t)) \quad \text{for all } t.$$

If $H_0(t) \equiv 0$, then $(R_z H_0)(t) \geqq H_0(t)$, and for each $_z$ and t, $R_z^j H_0(t)$ is a monotone increasing sequence in $j = 1, 2, \cdots$.

By assumptions (A1)-(A3), for all $t>0$ and all $z>0$,

$$
\begin{aligned}
(R_zH)(t) &\leqq \int_0^t DI^{-1}\left(\frac{1}{sD}\int_0^s G_1(T-v)I(G_2(T-v)(H(v)+z))\,dv\right) ds \\
\text{(A5)} \qquad &\leqq D\int_0^t I^{-1}\left(\frac{1}{KsD}\int_0^s I(K(H(v)+z))\,dv\right) ds \\
&\leqq D\int_0^t I^{-1}\left(\frac{k}{D} I(K(H(s)+z))\right) ds.
\end{aligned}
$$

But for each $z>0$, there exists by standard theorems and (A5) and (A1) with $C=KD$, a unique solution $\zeta(t,z)$ on $[0,T]$ of the equation

$$
\text{(A6)} \qquad \frac{d}{dt}\zeta(t)=DI^{-1}\left(\frac{k}{D}I(K(z+\zeta(t)))\right), \qquad \zeta(0)=0.
$$

The function $\zeta(\cdot)=\zeta(\cdot,z)$ evidently belongs to the domain $\mathscr{A}_T$ of R_z, and $H_0(t)\leqq(R_z\zeta)(t)$. In addition, $(R_z\zeta)(t)\leqq\zeta(t)$, by (A5) and (A6). Therefore $R_z^jH_0(t)\leqq\zeta(t)$ for all j and all $t\in[0,T]$, and $W_*(t)\equiv\lim_{j\to\infty}R_z^jH_0(t)$ exists and is an element of $\mathscr{A}_T$. Clearly, $W_*(t)=(R_zW_*)(t)$ is a minimal nonnegative solution of the integral fixed-point equation (A4′), and similarly $W^*(t)\equiv\lim_{j\to\infty}R_z^j\zeta(t)$ is a maximal solution since all solutions $W(t)$ of (A4′) are bounded by $\zeta(t)$, according to Theorem 4.1 of [5, pp. 25-26].

To show that the solution of (A4) or (A4′) is unique, we prove that the function $J(w)\equiv F(\int_t^T G_1(s)I(G_2(s)w)\,ds,t,T)$ is Lipschitz for $w\in[\delta',\infty)$ for each $\delta>0$, where $\delta'\equiv A_1kI^{-1}(k/\delta)$. Indeed, if $w\geqq v\geqq\delta'$, then both $I(v)$ and $I(w)$ are $\leqq\delta^{-1}$ by (A1.2), and by (A2)

$$
\begin{aligned}
|J(w)-J(v)|\leqq K_2\Bigg[&I^{-1}\left(\frac{1}{T-t}\int_t^T G_1(s)I(G_2(s)w)\,ds\right) \\
&-I^{-1}\left(\frac{1}{T-t}\int_t^T G_1(s)I(G_2(s)v)\,ds\right)\Bigg]
\end{aligned}
$$

for some constant K_2. By Lemma A.0

$$
\frac{1}{T-t}\int_t^T G_1(s)\begin{pmatrix} I(G_2(s)w) \\ I(G_2(s)v)\end{pmatrix} ds = G_1(s)\begin{pmatrix} I(G_2(s_2)w) \\ I(G_2(s_2)v)\end{pmatrix}
$$

for some values s_1 and s_2 in $[t,T]$ which depend on v and w. For fixed $t<T$, apply (A1) to deduce the Lipschitz property

$$
|J(w)-J(v)|\leqq K_2A(\delta)|w-v|.
$$

But all solution-functions W in $\mathscr{A}_T$ of the equation (A4′) are bounded between zero and the function $\zeta(\cdot,z)$ on $[0,T]$ by Theorem 4.1 of [5, pp. 25-26]. Thus all solutions W can be bounded uniformly on $[0,T]$ for each fixed $z>0$. It follows by standard arguments that the functions $W_*(t)$ and $W^*(t)$ coincide and that the solution $\underline{Z}(t,z)$ of (A4) or (A4′) is unique and depends continuously on z for $z\geqq\delta'$.

We know that $Z_1(0,z)\geqq z$ goes to ∞ as z does. Also, for each $z>0$, $Z_1(0,z)\leqq z+\zeta(T,z)$, and it is easy to see from (A6) and (A1.2) that $\zeta(T,0+)=0$. Thus $Z_1(0,z)$ goes to zero as z goes to zero. Therefore, the intermediate value theorem implies that for each $x>0$ there is a unique value of z for which the solution $\underline{Z}(\cdot,z)$ of (A4) satisfies $Z_2(0,z)=\int_0^T G_1(t)I(G_1(t)Z_1(t,z))\,dz=x$. Then $y_0=y_0(T,x)\equiv Z_1(0,z)$. Since $Z_1(t,z)$ increases with z for all t, it is easy to check that $Z_2(t,z)$ decreases with z, so that x decreases with increasing z, and y_0 is a decreasing function of x.

Next we check the upper bound for y_0 in terms of x/T. Since $Z_1(t, z) \leqq z + \zeta(T, z)$ for $0 \leqq t \leqq T$, we have for each s in $[0, T]$,

$$\frac{Z_2(s, z)}{T-s} = \frac{1}{T-s} \int_s^T G_1(t) I(G_2(t) Z_1(t, z))\, ds$$
$$\geqq \frac{1}{T-s} \int_0^T G_1(s) I(G_2(s)(z + \zeta(T, z)))\, ds \geqq kI(K(z + \zeta(T, z)))$$

by (A3) and the monotonic-decreasing property of $I(\cdot)$. Then by (A2),

$$y_0 = Z_1(0, z) = z + \int_0^T F(Z_2(s, z), s, T)\, ds$$
$$\leqq z + D \int_0^T I^{-1}\left(\frac{Z_2(s, z)}{(T-s)D}\right) ds$$
$$\leqq z + DTI^{-1}\left(\frac{k}{D} I(K(z + \zeta(T, z)))\right).$$

However, assumption (A1.2) for $C \geqq KD$ implies that

$$\frac{d}{dt} \zeta(t) \leqq DA_1(z + \zeta(t)), \qquad t \geqq 0$$

so that $\zeta(T, z) \leqq A_2 z$ for all $T \leqq T_*$ and some constant A_2 which will depend on T_*. Substituting in the upper bound for y_0, and applying (A1.2) once more, yields

$$y_0 \leqq (1 + DT_* A_1 A_2) z \equiv A_3 z.$$

Finally,

$$x = Z_2(0, z) = \int_0^T G_1(s) I(G_2(s) Z_1(s, z))\, ds \leqq KTI(kz) \leqq KTI(ky_0/A_3)$$

implies the desired upper bound $y_0 \leqq BI^{-1}(bx/T)$, where $B = A_3/k$, $b = 1/K$.

Clearly, y_0 is continuous with respect to x and T, and its Lipschitz dependence on x/T over compact intervals is obtained from (A2) and estimates like the preceding on the relations

$$y_0 = z + \int_0^T F(Z_2(t, z), t, T)\, dt, \qquad x = \int_0^T G_1(s) I(G_2(s) Z_1(s, z))\, ds.$$

The idea is to fix arbitrary $z, z' \geqq \delta'$, to define y_0' and x' from z' just as y_0 and x are defined from z, and then to bound difference quotients $(y_0' - y_0)/(z' - z)$ above and $(x' - x)/(T(z' - z))$ both above and below. Our proof is now complete. □

Acknowledgments. We are extremely grateful to Ioannis Karatzas for ideas that were very helpful in stimulating our proof of Lemma A.1 of the Appendix, and to the associate editor and an anonymous referee for suggesting numerous improvements in our presentation.

REFERENCES

[1] P. BREMAUD, *Point Processes and Queues: Martingale Dynamics*, Springer-Verlag, New York, 1981.

[2] J. C. COX AND C.-F. HUANG, *Optimal consumption and portfolio policies when asset prices follow a diffusion process*, J. Econom. Theory, 49 (1989), pp. 33–83.

[3] J. M. HARRISON AND S. R. PLISKA, *Martingales and stochastic integrals in the theory of continuous trading*, Stochastic Process. Appl., 11 (1981), pp. 215-260.

[4] ———, *A stochastic calculus model of continuous trading: complete markes*, Stochastic Process. Appl., 15 (1983), pp. 313-316.

[5] P. HARTMAN, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1973.

[6] H. HOTELLING, *The economics of exhaustible resources*, J. Political Economy, 39 (1931), pp. 137-175.

[7] I. KARATZAS, P. LAKNER, J. P. LEHOCZKY, AND S. E. SHREVE, *Dynamic equilibrium in a multi-agent economy: construction and uniqueness*, Proc. Nat. Acad. Sci., submitted.

[8] I. KARATZAS, J. P. LEHOCZKY, AND S. E. SHREVE, *Optimal portfolio and consumption decisions for a "small investor" on a finite horizon*, SIAM J. Control Optim., 25 (1987), pp. 1557-1586.

[9] P. LAKNER, *Optimal consumption and investment on a finite horizon with stochastic commodity prices*, in Linear Circuits, Systems and Signal Processing: Theory and Application, Elsevier Science Publishers B.V., North-Holland, 1988.

[10] ———, *Consumption/investment and equilibrium in the presence of several commodities*, Ph.D. thesis, Statistics Department, Columbia University, New York, 1989.

[11] R. C. MERTON, *Lifetime portfolio under uncertainty: the continuous-time case*, Rev. Econom. Statist., 51 (1969), pp. 247-257.

[12] ———, *Optimum consumption and portfolio rules in a continuous-time model*, J. Econom. Theory, 3 (1971), pp. 373-413.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 3, pp. 656-677, May 1991

009

THE WAVE METHOD FOR DETERMINING THE ASYMPTOTIC DAMPING RATES OF EIGENMODES I: THE WAVE EQUATION ON A RECTANGULAR OR CIRCULAR DOMAIN*

JIANXIN ZHOU† AND GOONG CHEN†

Abstract. The uniform exponential decay property of the wave equation with viscous boundary damping has been studied by several people. The mathematical proofs used therein are commonly based on energy identities, which *cannot determine the actual decay rates* of the solution. In Quinn and Russell [*Proc. Roy. Soc. Edinburgh Sect. A*, 77 (1977), pp. 97-127] and Chen [Ph.D. thesis, University of Wisconsin, Madison, WI, May 1977], such damping rates have been calculated for rectangular and circular domains, respectively, using separation of variables and a perturbation approach good for *small* viscous damping parameters. In this paper, we extend some earlier *geometrical optics* and diffraction methods of Keller and Rubinow [*Ann. Phys.*, 9 (1960), pp. 24-75] to treat the eigenvalue problems with dissipative conditions. Such methods provide strong insights into the physical properties of the solutions. Asymptotic estimates of damping and wavenumber are shown to agree favorably with the earlier results of Quinn and Russell and Chen for small damping parameters, as well as with the numerical solutions computed herein for cases even when the damping parameters are not small.

Key words. wave method, asymptotic damping rates, geometrical optics

AMS(MOS) subject classifications. 93D15, 93B60, 35L05, 35P20

1. Introduction.

Consider the wave equation

$$w_{tt}(x,t)-\Delta w(x,t)=0, \qquad x\in\Omega\subset\mathbb{R}^n, \quad t>0 \tag{1.1}$$

on a bounded domain Ω in $\mathbb{R}^n$. On the boundary $\partial\Omega$ of Ω, assume a dissipative condition

$$\frac{\partial w(x,t)}{\partial t}+\alpha\frac{\partial w(x,t)}{\partial n}=0, \qquad \alpha>0, \quad x\in\partial\Omega, \quad t>0, \tag{1.2}$$

where n is the unit outward normal on $\partial\Omega$. Condition (1.2) says that force $\partial w/\partial n$ is negatively proportional to the velocity, so a viscous damper is installed in effect everywhere on the boundary. We call α the viscous damping parameter. The initial conditions for the wave equation are

$$\left.\begin{aligned} w(x,0)&=w_0(x),\\ \frac{\partial w}{\partial t}(x,0)&=v_0(x),\end{aligned}\right\} \qquad x\in\Omega.$$

The energy of the system at time t is

$$E(t)=\int_\Omega [|\nabla w(x,t)|^2+w_t^2(x,t)]\,dx. \tag{1.3}$$

Several people (cf. [2], [8], [11], e.g.) have shown that if the domain Ω has certain geometries, then the energy of vibration will decay uniformly exponentially: there exist $M\geqq 1$, $\mu>0$ such that

$$E(t)\leqq Me^{-\mu t}E(0) \tag{1.4}$$

* Received by the editors July 24, 1989; accepted for publication (in revised form) May 29, 1990.

† Department of Mathematics, Texas A&M University, College Station, Texas 77843. This research was supported in part by Air Force Office of Scientific Research grant 87-0334.

for all initial conditions $(w_0, v_0) \in H^1(\Omega) \oplus H^0(\Omega)$, where $H^k(\Omega)$ is the Sobolev space of order $k \geqq 0$. This uniform decay property is important in the design of linear-quadratic regulators for distributed parameter control systems. The larger μ is, the higher the performance of the regulator is commonly regarded. Thus, from the practical design viewpoint, it is useful to know the overall damping rate μ or the damping rates of individual eigenmodes of the controlled system. Nevertheless, except in [1] and [11], explicit information on such damping rates is not available because the mathematical proofs as given in [2] and [8] are based on energy identities, which cannot provide sharp estimates of decay in relation to the damping parameter α. Indeed, such sharp estimates are pretty hard to obtain for general domains. Therefore, the best hope is to look for some answers in the case of special domains: rectangles, circles, and ellipses, etc.

Quinn and Russell [11] studied the case of rectangular geometry, while Chen [1] did the circular geometry. Both works are based on the elementary separation of variables approach and a perturbation argument valid for a *small damping parameter* α, and provide certain useful estimates for the damping rates of eigenmodes. It is obvious that the perturbation argument used therein is also valid when the *damping parameter* α *is large*—we simply develop a perturbation series in terms of $1/\alpha$ in lieu of α. But when α is of medium size—neither too large, nor too small—then the perturbation argument in [1] and [11] would fail and no *explicit* damping rates of eigenmodes would be available. A more *general* approach is required in order to obtain such information.

In this series of papers, we plan to generalize the *wave propagation method* (WPM) developed by us in [3] and [4], and elsewhere, to study the problem. The WPM in [4] gives sharp asymptotic estimates of eigenfrequencies of vibration for equations in *one* space dimension, including the wave, beam, and Schrödinger equations. The extension of WPM to higher space dimensions, say $n=2$ in (1.1), is difficult because more complicated wave phenomena such as caustics, foci, grazing and gliding rays can happen. Fortunately, for simple geometries such as rectangles, circles, and ellipses, the fundamental work has been done in an important paper [7], wherein Keller and Rubinow used the general geometrical theory of optics (GTO) and the geometrical theory of diffraction (GTD) to estimate eigenvalues of the Laplacian subject to Dirichlet and Neumann boundary conditions. Their methods and results are still unsurpassed to this day, even though the paper is already three decades old. In this paper, we will incorporate our earlier ideas of WPM [4] into Keller and Rubinow's work [7] to obtain asymptotic eigenfrequency estimates for the Laplacian with dissipative boundary conditions. Our results agree with those in [1] and [11] when the damping parameter α is small, and agree sharply with the numerical results when the damping parameter is not small. Our analysis here is based on the physical postulates of wave propagation and optics, which have all been rigorously justified in the asymptotic sense by Maslov and Fedoriuk [10].

Our plan is as follows. In the present paper, Part I, we will study mainly the *circular and rectangular* cases in *two* space dimension. In the follow-up paper, Part II, we plan to study the case of a rectangular *plate*, generalizing Keller and Rubinow's work to the *biharmonic* operator Δ^2. There are additional interesting problems such as:

(i) The cases of elliptical and triangular geometries;
(ii) Overdamped modes (cf. Remark 4.2); and
(iii) Space dimension three.

If possible, we also wish to investigate them in subsequent parts of this series of papers.

2. Extending Keller and Rubinow's wave method to dissipative boundary conditions. Let Ω be a bounded domain in $\mathbb{R}^3$ or $\mathbb{R}^2$. We let the damping parameter α in (1.2) be complex:

$$\frac{\partial w(x,t)}{\partial t}+\alpha\frac{\partial w(x,t)}{\partial n}=0, \qquad x\in\partial\Omega, \quad \alpha\in\mathbb{C}, \quad \Re\alpha>0 \tag{2.1}$$

($\Re\alpha$ and $\Im\alpha$ denote, respectively, the real and imaginary parts of α). This condition is more general than (1.2) in that it also covers some acoustic and electromagnetic scattering problems wherein the impedance parameter α is complex in general.

We consider an eigenmode $\phi(x)$:

$$w(x,t)=e^{-ikt}\phi(x), \qquad x\in\Omega. \tag{2.2}$$

Substituting (2.2) into (1.1) and (2.1), we obtain a reduced wave (i.e., Helmholtz) equation with impedance boundary condition:

$$\begin{aligned}&(\Delta+k^2)\phi(x)=0, \qquad x\in\Omega,\\ &ik\phi(x)-\alpha\frac{\partial\phi(x)}{\partial n}=0, \qquad x\in\partial\Omega.\end{aligned} \tag{2.3}$$

Our task here is to derive estimates on the eigenfrequencies k.

To make our paper sufficiently self-contained, we first introduce the theory developed by Keller and Rubinow in [7], with suitable adaptations to our problem. (Not surprisingly, many passages and figures in this section are adopted from [7] without alteration.)

According to GTO [5], [6], when $|k|$ is large, ϕ can be represented by a geometrical optics expansion:

$$\begin{aligned}\phi(x)&=\sum_{j=1}^{N}e^{ikS_j(x)}\left[\sum_{p=0}^{\infty}\frac{A_{j,p}(x)}{(ik)^p}\right], \qquad x\in\Omega\\ &=\sum_{j=1}^{N}e^{ikS_j(x)}\left[A_{j,0}(x)+O\left(\frac{1}{k}\right)\right],\end{aligned} \tag{2.4}$$

where it is assumed there are N waves propagating in Ω. Each wave has phase S_j and (dominant) amplitude $A_{j,0}$. For simplicity, let us write $A_{j,0}$ as A_j. Substituting (2.4) into (2.3) and equating to zero the coefficients of k^2 and k, we obtain

$$|\nabla S_j|^2=1 \quad \text{(the eiconal equation)}, \tag{2.5}$$

$$2\nabla S_j\cdot\nabla A_j+A_j\Delta S_j=0 \quad \text{(the transport equation)} \tag{2.6}$$

for $j=1,2,\cdots,N$.

The surfaces $S_j=\text{constant}$ are called wavefronts, and the trajectories orthogonal to the wavefronts are called rays. The rays, as characteristic to (2.5), are easily seen to be straight lines. Let τ denote the arclength along a ray; then (2.6) implies that along the ray S_j is given by

$$S_j(\tau)=S_{j0}\pm\tau, \tag{2.7}$$

where S_{j0} is the value of S_j at the point from which τ is measured. The sign "±" in (2.7) can be chosen to be "+" or "−" depending on if τ is measured positively or negatively in the direction of increasing S_j.

The transport equation (2.6) is a linear ordinary differential equation along rays

$$2\frac{dA_j}{d\tau}+A_j(\Delta S_j)=0, \qquad j=1,2,\cdots,N,$$

which has solution

$$A_j(\tau)=A_j(0)\exp\left(-\frac{1}{2}\int_0^\tau \Delta S_j(\eta)\,d\eta\right)=A_j(0)\left[\frac{G_j(\tau)}{G_j(0)}\right]^{1/2}. \tag{2.8}$$

$G_j(\tau)$ is the Gaussian curvature of the wavefront $S_j=\text{const.}$ at τ. When $n=3$, (2.8) is further seen to be equal to

$$A_j(\tau)=A_j(0)\left[\frac{\rho_{j1}\rho_{j2}}{(\rho_{j1}+\tau)(\rho_{j2}+\tau)}\right]^{1/2} \quad (\text{in } \mathbb{R}^3), \tag{2.9}$$

where ρ_{j1} and ρ_{j2} denote the principal radii of curvature of the wave front at $\tau=0$. When $n=2$, we regard $\rho_{j2}=\infty$ in (2.9) (i.e., a cylindrical wavefront) so we can write

$$A_j(\tau)=A_j(0)\left(\frac{\rho_j}{\rho_j+\tau}\right)^{1/2} \quad (\text{in } \mathbb{R}^2). \tag{2.10}$$

We now apply the boundary conditions in (2.3) to the solution (2.4). We substitute and equate to zero the coefficient of k to obtain

$$\sum_{j=1}^{N}\left(A_j-\alpha A_j\frac{\partial S_j}{\partial n}\right)e^{ikS_j}=0. \tag{2.11}$$

As in [7], we now assume that at every point on the boundary the terms in (2.11) for which $A_j-\alpha A_j\,\partial S_j/\partial n\neq 0$ *vanish in pairs*. That is to say, for each such j wave, there is another j' wave, $j'\neq j$, such that

$$\left(A_j-\alpha A_j\frac{\partial S_j}{\partial n}\right)e^{ikS_j}+\left(A_{j'}-\alpha A_{j'}\frac{\partial S_{j'}}{\partial n}\right)e^{ikS_{j'}}=0 \quad \text{on } \Omega. \tag{2.12}$$

Physically, this assumption means that each wave hitting the boundary gives rise to a reflected wave. (The waves for which $A_j-\alpha A_j\,\partial S_j/\partial n=0$ are absorbed by the boundary and do not give rise to reflected waves.) Since (2.12) holds for a range of values of k, it follows that

$$S_j=S_{j'} \quad \text{on } \partial\Omega. \tag{2.13}$$

Therefore,

$$\frac{\partial S_j}{\partial s}=\frac{\partial S_{j'}}{\partial s} \quad \text{on } \partial\Omega, \tag{2.14}$$

where $s=(-n_2,n_1)$ is the unit tangent vector in the counterclockwise sense, with (n_1,n_2) being the components of the unit outward normal n. Since

$$\nabla S_k=\frac{\partial S_k}{\partial s}s+\frac{\partial S_k}{\partial n}n \quad \text{for } k=1,2,\cdots,N,$$

from (2.5) and (2.14), we obtain

$$\left\|\frac{\partial S_j}{\partial n}n\right\|^2=\left\|\frac{\partial S_{j'}}{\partial n}n\right\|^2.$$

Therefore,

$$\frac{\partial S_j}{\partial n}=\pm\frac{\partial S_{j'}}{\partial n} \quad \text{on } \partial\Omega. \tag{2.15}$$

Arguing along the same lines as in [7], we can show that the "+" sign in (2.15) is not admissible; thus,

$$\frac{\partial S_j}{\partial n} = -\frac{\partial S_{j'}}{\partial n} \quad \text{on } \partial\Omega. \tag{2.16}$$

Using (2.13) and (2.16), we obtain

$$A_{j'} = -\frac{1-\alpha(\partial S_j/\partial n)}{1+\alpha(\partial S_j/\partial n)} A_j \quad \text{on } \partial\Omega. \tag{2.17}$$

We call this the *reflection relation.* Note that if we choose $\alpha = 0$ and $\alpha = \infty$, then we obtain, respectively, the reflection relations for the Dirichlet and Neumann boundary conditions in [7].

Equation (2.9) for the amplitude A_j breaks down at points on a *caustic surface,* where $\tau = -\rho_{j1}$ or $\tau = -\rho_{j2}$ ($\rho_{j1} \neq \rho_{j2}$). (If $\tau = -\rho_{j1} = -\rho_{j2}$, such a point is called a *focal* point.) Indeed, near a caustic surface a more elaborate WKB approximation is necessary for a uniformly valid expansion of the field (cf. the book by Maslov and Fedoriuk [10] for a general treatment of wave optics on manifolds; cf. also Remark 4.2).

As in [7], we assume that

(i) Each wave converging to a caustic gives rise to another wave diverging from the caustic; and

(ii) The rays of the diverging wave (which are *gliding rays*) are assumed to be the continuations of those of the converging wave and the phase along these rays is assumed to be the continuation of the phase on the converging rays.

Thus,

$$S_j = S_{j'} \tag{2.18}$$

on a caustic. By passing a regular (i.e., nonfocal) point on a caustic, in (2.9), $\rho_{j1} + \tau$ (or $\rho_{j2} + \tau$) changes sign from "−" to "+." Thus

$$A_{j'} = e^{-i\pi/2} A_j. \tag{2.19}$$

(Passing a focal point, the relation is

$$A_{j'} = e^{-i\pi} A_j \tag{2.20}$$

because both $\rho_{j1} + \tau$ and $\rho_{j2} + \tau$ change signs from "−" to "+.")

Now let us trace a ray of any wave in the direction of increasing τ. We come to a caustic or boundary because Ω is a bounded domain. In either case, the ray continues as a ray of another wave. A sequence of waves is encountered in this manner. Since there are, by assumption, only a finite number N of waves in the solution, one of the waves in this sequence must *recur.* Therefore, a ray orthogonal to a given wavefront is ultimately orthogonal to this very same wave front again. But the value of S continually increases as a ray is traversed in the positive direction. Therefore, at the second point of intersection of the wavefront and the ray, the value of S_j is greater than its initial value by the length of the ray between intersections. Since S_j is constant on a wavefront, S_j must therefore be *multiple valued.* The corresponding amplitude A_j may also be *multiple valued* due to the change (2.19) or (2.20) on the ray paths.

To form an eigenmode ϕ, the solution to (2.3), *resonance must occur*; i.e., the wave that recurs must be *in phase with itself,* that is to say, the overall phase difference due to the changes of S_j and A_j must be an integral multiple of 2π. This is expressed by

$$k(\delta S_j) + \delta(\arg A_j) = 2\pi n_j, \qquad j = 1, \cdots, N, \ n_j \in \mathbb{Z}, \tag{2.21}$$

where δS_j and $\delta(\arg A_j)$ denote, respectively, the changes of S_j and the argument of the complex amplitude A_j.

In [7], Keller and Rubinow further used the important concept of a *covering space* to resolve the difficulty of the multivaluedness of S_j by regarding them as branches of a single function S. The covering space is *multisheeted*, just like a Riemann surface in complex analysis. The number of sheets is equal to the N distinct branches of ∇S. The various sheets are replicas of the domain Ω which may be bounded internally by caustics. The sheets corresponding to ∇S_{j1} and ∇S_{j2} are joined together along the part of the caustic or boundary where $S_{j1}=S_{j2}$, the places where the wave j_1 gives rise to the wave j_2 by reflection or by passing through a caustic. Similarly, we consider A_j to be branches of a single function A defined on the covering space.

Now assume that the fundamental group of the covering space (of ∇S) contains q linearly independent closed curves $C_j, j=1,2,\cdots,q$. Then the condition (2.21) can be rewritten as

$$k\oint_{C_j} \nabla S\cdot d\sigma+\sum \delta(\arg A)=2\pi n_j, \qquad j=1,2,\cdots,q. \tag{2.22}$$

We will incorporate the reflection condition (2.17) into the formula (2.22) to obtain the asymptotic damping rates and eigenfrequency estimates in the examples to be discussed in the next few sections.

3. Asymptotic damping rates and eigenfrequencies for a circular domain. Circular and elliptical domains seem to be the only domains where caustics are known to be constructible [7]. We consider a $(2-D)$ disk Ω with radius a. Let $C(a_0)$ be a circle with radius a_0 and centered at the origin. Any ray in Ω tangent to $C(a_0)$ will be tangential to it again after a reflection at $\partial\Omega$ (cf. Fig. 1). Thus $C(a_0)$ is a caustic to the family of rays generated by successive reflections at $\partial\Omega$.

To apply Keller and Rubinow's theory in § 2, we need to find a set of N normal congruences of rays that are closed under reflection. (A normal congruence of rays is a family of rays orthogonal to any given surface.) From the preceding paragraph, we can consider all those rays traveling inward from $\partial\Omega$ to the caustic $C(a_0)$ as one normal congruence (cf. Fig. 2) and all those traveling outward from $C(a_0)$ to $\partial\Omega$ as a second

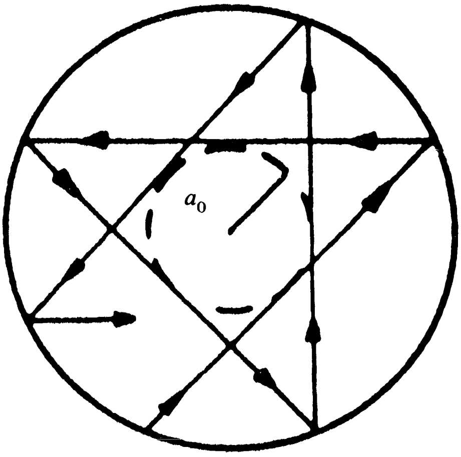


FIG. 1. *A ray inside a circular domain and some of the rays that arise from it after several reflections. All of these rays are tangential to a concentric circle of radius* a_0, *the caustic.* (*Reprinted from* [7], *with permission.*)

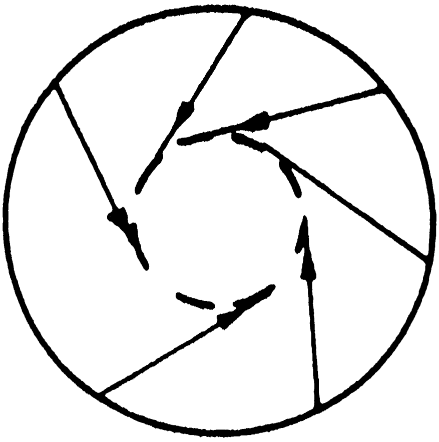


FIG. 2. *Normal congruence* I *of rays converging to the caustic.* (*Reprinted from* [7], *with permission.*)

congruence (cf. Fig. 3). Therefore, $N = 2$. The covering space for a disk is topologically homeomorphic to a torus [7]. There are only two linearly independent closed curves on a torus, so $q = 2$ in (2.22). From the way the covering space is constructed in [7], we can choose the first curve C_1 to be just $C(a_0)$, and choose C_2 to be shown as in Fig. 4.

We now apply (2.22) to C_1 and C_2. Along C_1, ∇S is tangent to it, and no ray crosses the caustic, so

$$\oint_{C(a_0)} \nabla S \cdot d\sigma = \text{arclength of } C(a_0) = 2\pi a_0,$$

$$\delta(\arg A) = 0,$$

and (2.22) gives

$$k(2\pi a_0) = 2\pi l, \qquad l = 0, 1, 2, \cdots. \tag{3.1}$$

Remark 3.1. It is well understood that for the eigenvalue problem (2.3), most of the true eigenvalues $\lambda = -k^2$ are *complex numbers.* This implies that (most of) the k themselves are complex numbers. So how can a complex number be equal to a real number in (3.1)? It turns out that k *in* (3.1) *should be regarded as a real number*, namely,

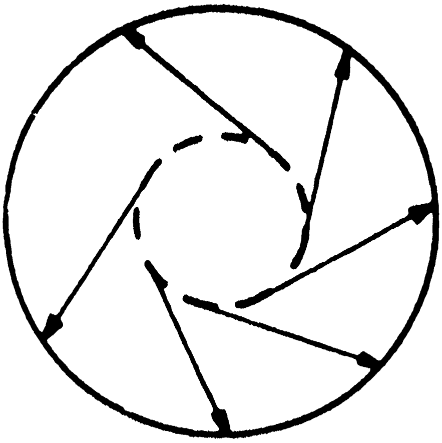


FIG. 3. *Normal congruence* II *of rays diverging from the caustic.* (*Reprinted from* [7], *with permission.*)

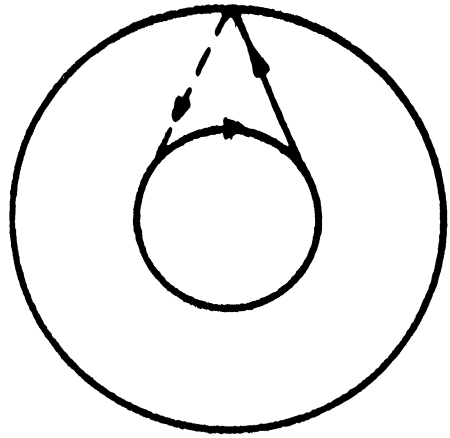


FIG. 4. *The closed path C_2 on the toroidal covering space.* (*Reprinted from* [7], *with permission.*)

the real part of the true eigenfrequence λ in Theorem 1 *later,* as k here is only an approximation to the real part of the true eigenfrequence λ. We can aptly call k the *wavenumber.* The rationale will become clear later on (see especially Remark 3.2 afterwards).

The curve C_2, as shown in Fig. 4, consists of two rays, each of length $(a^2-a_0^2)^{1/2}$, and an arc of the caustic of length $2a_0\cos^{-1}(a_0/a)$. Therefore,

$$\oint \nabla S\cdot d\sigma = 2\left(\sqrt{a^2-a_0^2}-a_0\cos^{-1}\frac{a_0}{a}\right). \tag{3.2}$$

Note the choice of "$-$" sign to conform to the counterclockwise sense on $C(a_0)$. Consider the change of arg A. This path crosses the caustic only once, so by (2.19) arg A is retarded by $\pi/2$. Also at the boundary, reflection occurs according to (2.17). Therefore, arg A changes by another angle

$$-\pi+\arg\frac{1-\alpha(\partial S/\partial n)}{1+\alpha(\partial S/\partial n)}. \tag{3.3}$$

Let us combine both changes of $\delta(\arg A)$. In consistency with the sign convention in [7], we actually have

$$\begin{aligned}\delta(\arg A)&=\left(-\frac{\pi}{2}\right)-\left[-\pi+\arg\left(\frac{1-\alpha(\partial S/\partial n)}{1+\alpha(\partial S/\partial n)}\right)\right]\\&=\frac{\pi}{2}-\arg\left(\frac{1-\alpha(\partial S/\partial n)}{1+\alpha(\partial S/\partial n)}\right)\end{aligned} \tag{3.4}$$

and using (3.1) in (3.2), we obtain

$$\begin{aligned}2k\left(\sqrt{a^2-a_0^2}-a_0\cos^{-1}\frac{a_0}{a}\right)&=2\pi m-\frac{\pi}{2}+\arg\left(\frac{1-\alpha(\partial S/\partial n)}{1+\alpha(\partial S/\partial n)}\right),\\ k\left(\sqrt{a^2-a_0^2}-a_0\cos^{-1}\frac{a_0}{a}\right)&=\pi\left[m-\frac{1}{4}+\frac{1}{2\pi}\arg\left(\frac{1-\alpha(\partial S/\partial n)}{1+\alpha(\partial S/\partial n)}\right)\right],\end{aligned} \tag{3.5}$$

$$m=1,2,3,\cdots.$$

We further simplify the above by noting from Fig. 5 (or cf. (32) of [7, p. 35]) that

$$S_1(r,\theta)=a_0\left[\theta-\cos^{-1}\left(\frac{a_0}{r}\right)\right]+(r^2-a_0^2)^{1/2}.$$

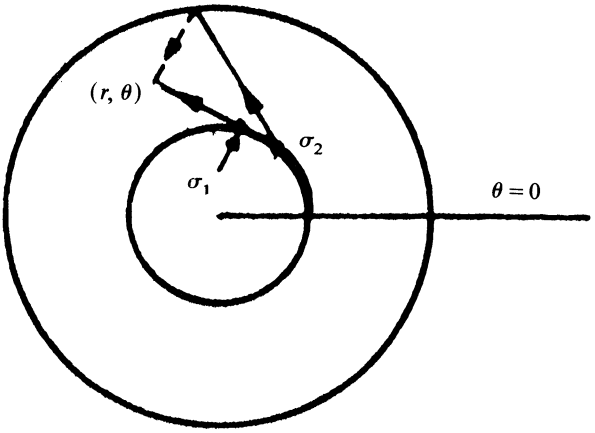


FIG. 5. *The two-ray paths from the caustic to the point* (r, θ). *One ray leaves the caustic at* $\sigma_1 = a_0[\theta - \cos^{-1})]$ *and travels a distance* $\tau_1 = (r^2 - a_0^2)^{1/2}$ *to the point. The other ray leaves the caustic at* $\sigma_2 = a_0[\theta + \cos^{-1}(a_0/r) - 2\cos^{-1}(a_0/a)]$, *is reflected from the boundary, and reaches the same point on the second sheet after traversing a distance* $\tau_2 = 2(a^2 - a_0^2)^{1/2} - (r^2 - a_0^2)^{1/2}$. (*Reprinted from* [7], *with permission.*)

Thus,

$$\left.\frac{\partial S_1}{\partial n}\right|_{\partial\Omega} = \left.\frac{\partial S_1}{\partial r}\right|_{r=a} = \frac{\sqrt{a^2 - a_0^2}}{a}$$

and (3.5) becomes

$$(3.6)\qquad k\left(\sqrt{a^2 - a_0^2} - a_0 \cos^{-1}\frac{a_0}{a}\right) = \pi\left[m + \frac{3}{4} + \frac{1}{2\pi}\arg\left(\frac{1 - \alpha\sqrt{a^2 - a_0^2}/a}{1 + \alpha\sqrt{a^2 - a_0^2}/a}\right)\right]$$

$$m = 0, 1, 2, \cdots.$$

Note that we have shifted m by one from (3.5).

Substituting a_0 from (3.1) into (3.6), we obtain

$$(3.7)\qquad [(ka)^2 - l^2]^{1/2} - l\cos^{-1}\left(\frac{l}{ka}\right) = \pi\left[m + \frac{3}{4} + \frac{1}{2\pi}\arg\left(\frac{1 - \alpha\sqrt{k^2a^2 - l^2}/ak}{1 + \alpha\sqrt{k^2a^2 - l^2}/ak}\right)\right]$$

$$m, l = 0, 1, 2, 3, \cdots.$$

This relation determines k, the (real) wavenumber, up to $O(1/k)$. Note that by choosing $\alpha = 0$ and $\alpha = +\infty$ in (3.6), we recover, respectively, formulas (26) and (25) of [7, pp. 33-34] for the Dirichlet and Neumann boundary conditions.

Next, we determine the damping rates of eigenmodes. We use the same idea as developed in [4]. The damping of wave motion is completely attributed to the *attenuation of the wave amplitude after reflection* at the boundary. According to the reflection relation (2.17), the amount of damping endured after each reflection is

$$\ln\left|-\frac{1 - \alpha(\partial S/\partial n)}{1 + \alpha(\partial S/\partial n)}\right| = \ln\left|\frac{1 - (\alpha/a)\sqrt{a^2 - a_0^2}}{1 + (\alpha/a)\sqrt{a^2 - a_0^2}}\right|.$$

Consider a *complete cycle* of wave motion; i.e., let a wave of normal congruence II reflect (at $\partial\Omega$) into a wave of normal congruence I, and then further transmit across the caustic to become a wave of normal congruence II (cf. Figs. 1, 2). Since the caustic is essentially a barrier against wave penetration, it can be regarded as a boundary surface (curve) also. As the wave propagates with speed one, and the round trip distance made by the wave from the circumference to the caustic is $2(a^2-a_0^2)^{1/2}$ (cf. Fig. 4) by using the very same arguments as in the one-dimensional wave propagation case in [4, § 2], we obtain

$$\text{rate of damping per unit time} = \frac{\text{total attenuation in a cycle of wave motion}}{\text{the round trip time duration}},$$

i.e.,

$$\mu(k) = \frac{1}{2\sqrt{a^2-a_0^2}} \ln \left| \frac{1-(\alpha/a)\sqrt{a^2-a_0^2}}{1+(\alpha/a)\sqrt{a^2-a_0^2}} \right| \left(= \frac{1}{2\sqrt{a^2-(l/k)^2}} \ln \left| \frac{ak-\alpha\sqrt{k^2a^2-l^2}}{ak+\alpha\sqrt{k^2a^2-l^2}} \right| \right). \tag{3.8}$$

We can now state Theorem 1.

THEOREM 1. *Let $\lambda \in \mathbb{C}$, $\Im\lambda \leqq 0$, satisfy*

$$\begin{aligned} &(\Delta+\lambda^2)\phi(x)=0, && x\in\Omega = \textit{the disk with radius } a,\\ &i\lambda\phi(x)-\alpha\frac{\partial\phi(x)}{\partial n}=0 && \textit{on } \partial\Omega. \end{aligned} \tag{3.9}$$

Then $k+i\mu(k)$ is an approximation to some λ such that

$$\lambda = k+i\mu(k)+O\left(\frac{1}{k}\right) \quad \textit{for large } k\in\mathbb{R}, \tag{3.10}$$

where k and $\mu(k)$ satisfy, respectively, (3.7) *and* (3.8) *for given integers l, $m=0,1,2,\cdots$.*

Remark 3.2. Our treatment in this section is based on the right *physics* of the problem. As the waves are propagating in the medium Ω, there is no loss of energy; therefore, *the wavenumber k should be real,* at least asymptotically. The only energy loss happens when the wave hits the boundary and makes the reflection. Therefore, nearly the entire damping rate can be attributed to the decrease of magnitude of the amplitude.

4. Comparison with existing results and numerical solutions. The case where Ω is the *unit disk* has been treated by Chen in his Ph.D. thesis [1]. Since those results have never been published elsewhere, we briefly summarize them below.

A direct approach of separation of variables in polar coordinates

$$\phi(x)=J_l(\lambda r)\,e^{\pm il\theta}, \quad J_l\text{: the Bessel function of order } l \tag{4.1}$$

for (3.9) leads to the transcendental equation

$$J_l(\lambda)+i\alpha J_l'(\lambda)=0, \qquad l=0,1,2,\cdots, \tag{4.2}$$

for the determination of the eigenfrequency λ. Let λ_{lm} denote the mth positive zero of the Bessel function J_l. We divide our discussion into the following cases:

(a) *Small α.* It has been shown in [1] that each λ_{lm} has a small neighborhood (uniform with respect to l, m) such that the solution λ of (4.2) is analytic with respect *to the damping parameter* α:

$$(4.3)\qquad \lambda = \lambda_{lm} + C_{lm1}\alpha + C_{lm2}\alpha^2 + \cdots + C_{lmj}\alpha^j + \cdots, \quad |\alpha| \text{ sufficiently small.}$$

We can actually compute a few coefficients by substituting (4.3) into (4.2) and using the Taylor expansion:

$$C_{lm1} = -i,$$

$$C_{lm2} = -\frac{3}{2}\frac{1}{\lambda_{lm}},$$

$$\vdots$$

(cf. [1, p. 114]). Thus for small α, the rate of damping is approximately equal to

$$(4.4)\qquad \Re(-i\lambda) = -\Re\alpha + O(\alpha^2).$$

We have computed a set of zeros of (4.2) for small α by using Newton's method, for $l = 0, 1, 4$. See Table 1. Using $\alpha = 0.01$, 0.015, 0.02, respectively, we have found from Table 1 that the damping rates are approximately equal to those in Table 2, in close agreement with (4.4).

To make comparisons, now we apply our formula (3.8) to the unit disk.

COROLLARY 1. *Consider the unit disk where* $a = 1$. *Then for* $\alpha \in \mathbb{C}$ *such that* $|\alpha|$ *is small and* $\Re\alpha > 0$, *we have*

$$\mu(k) = -\Re\alpha + O(\alpha^2).$$

Proof. Write

$$\alpha = \eta + i\xi, \qquad \eta, \xi \in \mathbb{R}.$$

We note that the caustic radius a_0 satisfies $0 < a_0 < 1$. By (3.8),

$$\begin{aligned}\mu(k) &= \frac{1}{2\sqrt{1-a_0^2}} \ln\left|\frac{[1-\eta\sqrt{1-a_0^2}]^2 + (\xi\sqrt{1-a_0^2})^2}{[1+\eta\sqrt{1-a_0^2}]^2 + (\xi\sqrt{1-a_0^2})^2}\right| \cdot \frac{1}{2} \\ &= \frac{1}{4\sqrt{1-a_0^2}} \ln\left|\frac{1-2\eta\sqrt{1-a_0^2} + |\alpha|^2\sqrt{1-a_0^2}}{1+2\eta\sqrt{1-a_0^2} + |\alpha|^2\sqrt{1-a_0^2}}\right| \\ &= \frac{1}{4\sqrt{1-a_0^2}}[-4\eta\sqrt{1-a_0^2} + O(|\alpha|^2)] \\ &= -\eta + O(|\alpha|^2) \\ &= -\Re\alpha + O(|\alpha|^2). \qquad \square\end{aligned}$$

Thus Corollary 1 is in full agreement with (4.4).

(b) *Medium size* α, $\alpha \neq 1$. Again, we use Newton's method to compute the zeros of (4.2), using $\alpha = 0.2$, 0.5, $0.6 + i \cdot 0.3$, 0.95, 1.5. Here we only tabulate the case $l = 4$. (For other integral values of l, the numerical data have manifested the same pattern.) Refer to Table 3.

TABLE 1

Zeros of $J_l(\lambda)+i\alpha J_l'(\lambda)=0$ for $l=0,1,4$ and small α. The number on the left denotes the real part of λ, while the number on the right denotes the imaginary part of λ.

m	$l=0$	$l=1$	$l=4$
	$\alpha=0.0100$	$\alpha=0.0100$	$\alpha=0.0100$
1	2.404846, −0.010000	3.831719, −0.010000	7.588349, −0.010000
2	5.520087, −0.010000	7.015594, −0.010000	11.064714, −0.010000
3	8.653734, −0.010000	10.173473, −0.010000	14.372540, −0.010000
4	11.791539, −0.010000	13.323696, −0.010000	17.615969, −0.010000
5	14.930921, −0.010000	16.470633, −0.010000	20.826935, −0.010000
6	18.071067, −0.010000	19.615861, −0.010000	24.019022, −0.010000
7	21.211639, −0.010000	22.760087, −0.010000	27.199090, −0.010000
8	24.352474, −0.010000	25.903674, −0.010000	30.371009, −0.010000
9	27.493481, −0.010000	29.046830, −0.010000	33.537139, −0.010000
10	30.634608, −0.010000	32.189681, −0.010000	36.699002, −0.010000
	$\alpha=0.0150$	$\alpha=0.0150$	$\alpha=0.0150$
1	2.404872, −0.015001	3.831735, −0.015001	7.588357, −0.015001
2	5.520098, −0.015001	7.015602, −0.015001	11.064720, −0.015001
3	8.653741, −0.015001	10.173479, −0.015001	14.372545, −0.015001
4	11.791544, −0.015001	13.323700, −0.015001	17.615972, −0.015001
5	14.930925, −0.015001	16.470637, −0.015001	20.826938, −0.015001
6	18.071070, −0.015001	19.615864, −0.015001	24.019024, −0.015001
7	21.211642, −0.015001	22.760089, −0.015001	27.199092, −0.015001
8	24.352476, −0.015001	25.903676, −0.015001	30.371011, −0.015001
9	27.493483, −0.015001	29.046832, −0.015001	33.537141, −0.015001
10	30.634610, −0.015001	32.189683, −0.015001	36.699004, −0.015001
	$\alpha=0.0200$	$\alpha=0.0200$	$\alpha=0.0200$
1	2.404909, −0.020002	3.831758, −0.020002	7.588369, −0.020002
2	5.520114, −0.020002	7.015615, −0.020003	11.064728, −0.020002
3	8.653751, −0.020003	10.173488, −0.020003	14.372551, −0.020002
4	11.791551, −0.020003	13.323707, −0.020003	17.615977, −0.020003
5	14.930931, −0.020003	16.470642, −0.020003	20.826943, −0.020003
6	18.071075, −0.020003	19.615869, −0.020003	24.019028, −0.020003
7	21.211646, −0.020003	22.760093, −0.020003	27.199095, −0.020003
8	24.352480, −0.020003	25.903680, −0.020003	30.371014, −0.020003
9	27.493486, −0.020003	29.046835, −0.020003	33.537144, −0.020003
10	30.634613, −0.020003	32.189686, −0.020003	36.699007, −0.020003

TABLE 2

Damping rates for $l=0,1,3$ and small α

	$\alpha=0.001$	$\alpha=0.015$	$\alpha=0.02$
$l=0$	−0.010000	−0.015001	−0.020002 ~ 0.020003
$l=1$	−0.010000	−0.015001	−0.020002 ~ 0.020003
$l=4$	−1.010000	−0.015001	−0.020002 ~ 0.020003

TABLE 3
Zeros of $J_4(\lambda)+i\alpha J_4'(\lambda)=0$ *for medium size* α.

m	$\alpha=0.2$	$\alpha=0.5$	$\alpha=0.6+i\cdot 0.3$
1	7.591022, −0.201866	7.606712, −0.532225	7.226642, −0.628858
2	11.066571, −0.202322	11.078529, −0.540915	10.675812, −0.618721
3	14.373976, −0.202489	14.383550, −0.544249	13.973686, −0.612093
4	17.617143, −0.202570	17.625107, −0.545912	17.212268, −0.607839
5	20.827930, −0.202616	20.834740, −0.546866	20.420470, −0.604931
6	24.019885, −0.202645	24.025830, −0.547466	23.610813, −0.602830
7	27.199852, −0.202664	27.205127, −0.547868	26.789701, −0.601247
8	30.371693, −0.202678	30.376432, −0.548151	29.960780, −0.600013
9	33.537758, −0.202688	33.542060, −0.548358	33.126286, −0.599025
10	36.699568, −0.202695	36.703506, −0.548514	36.287671, −0.598217

m	$\alpha=0.95$	$\alpha=1.5$
1	7.686725, −1.306102	5.885229, −1.791908
2	11.201637, −1.476234	9.244764, −1.050649
3	14.521295, −1.572608	12.649875, −0.919150
4	17.764806, −1.634592	15.938485, −0.873084
5	20.970697, −1.677146	19.174790, −0.850710
6	24.155629, −1.707628	22.382899, −0.837951
7	27.327934, −1.730159	25.573933, −0.829925
8	30.492176, −1.747237	28.753790, −0.824525
9	33.651029, −1.760458	31.925909, −0.820708
10	36.806140, −1.770878	35.092441, −0.817906

We now compare the damping rates in Table 3 with the values obtained by applying formula (3.8). Using $a=1$, and noting that the radius a_0 of the caustic $C(a_0)$ shrinks to 0 as m becomes large, we obtain

(i) For $\alpha=0.2$

$$\frac{1}{2}\ln\left|\frac{1-0.2}{1+.02}\right|=-0.202733.$$

This is very close to the value −0.202695 at the bottom of the second column in Table 3.

(ii) For $\alpha=0.5$

$$\frac{1}{2}\ln\left|\frac{1-0.5}{1+0.5}\right|=-0.549306.$$

This is very close to the value −0.548514 at the bottom of the fourth column in Table 3.

(iii) For $\alpha=0.6+i0.3$,

$$\frac{1}{2}\ln\left|\frac{1-(06.+i\cdot 0.3)}{1+(0.6+i\cdot 0.3)}\right|=-0.581575,$$

which is close to −0.598217 at the bottom of the sixth column in Table 3. Note that the values in that column are increasing, contrasting the decreasing columns 2 and 4 in (i) and (ii).

(iv) For $\alpha = 0.95$,

$$\frac{1}{2}\ln\left|\frac{1-0.95}{1+0.95}\right| = -1.831781. \tag{4.5}$$

This deviates a little from the value 1.770878 at the bottom of column 8 in Table 3. The reason is that the damping parameter 0.95 is rather close to one. There is little question that those damping rates will still converge to the value in (4.5) if we let m become larger and larger.

(v) For $\alpha = 1.5$,

$$\frac{1}{2}\ln\left|\frac{1-1.5}{1+1.5}\right| = -0.804719. \tag{4.6}$$

(iv) This is close to the value -0.815787 at the bottom of column 10 in Table 3. We have checked that if more and more roots are computed (i.e., m is chosen large), then those damping rates will keep increasing and getting closer and closer to the value (4.6).

(c) *The characteristic impedance case*, $\alpha = 1$. When $\alpha = 1$, formula (3.8) gives

$$\frac{1}{2}\ln\left|\frac{1-1}{1+1}\right| = -\infty.$$

This infinitely large damping rate means that the amplitude of the reflected wave is zero; i.e., the incoming wave (here, the normal congruence II wave) would be *completely absorbed* by the boundary. In the *one-dimensional* case, such a complete wave absorption property is easy to confirm (see [3], e.g.; cf. also [9] for the higher-dimensional case). Let us look at the two-dimensional case here.

We have computed a large number of zeros of (4.2) for $l = 4$ again. These zeros are listed in Table 4. The reader can find that the rates of damping are steadily decreasing. For the 185th root, the damping rate is -3.881022. Since

$$2 \cdot e^{-3.881022} = 0.041259,$$

this means that at the wavenumber $k = 587.4475$ (i.e., the real part of the 185th zero of (4.2), when $l = 4$), the reflected wave (i.e., normal congruence I) has an amplitude that is only about 4.1 percent of the incoming wave.

We conjecture that the imaginary parts of the zeros of (4.2) for each l will decrease without bound, reaching $-\infty$.

From the stabilizer design point of view, choosing $\alpha = 1$ seems to suppress vibration most effectively.

We have observed that the damping rates listed in certain columns of various tables are steadily decreasing (see, e.g., columns 2, 4, and 8 in Table 3), whereas those listed elsewhere (e.g., columns 6 and 10 in Table 3) are steadily increasing.

This can be interpreted by Corollary 2.

COROLLARY 2. *Let $l \in \mathbb{Z}^+$ be fixed in* (3.8). *Then the rates of damping are decreasing if*

$$\frac{d}{da_0}\mu(k, a_0) < 0 \tag{4.7}$$

and increasing if

$$\frac{d}{da_0}\mu(k, a_0) > 0, \tag{4.8}$$

TABLE 4

Zeros of $J_4(\lambda)+i\alpha J_4(\lambda)=0$ *when* $\alpha=1$. *Note that the imaginary parts of* λ *all have negative signs as indicated by* $-\mathrm{Im}$.

	Re	−Im		Re	−Im		Re	−Im
	7.705903,	1.468978		204.117590,	3.352118		398.937859,	3.687488
	11.283858,	1.743047		207.260466,	3.359772		402.079794,	3.691412
	14.670407,	1.923818		210.403304,	3.367311		405.221723,	3.695305
	17.975551,	2.056575		213.546105,	3.374737		408.363648,	3.699168
	21.235540,	2.160406		216.688872,	3.382054		411.505567,	3.703002
	24.467162,	2.245244		219.831605,	3.389266		414.647481,	3.706806
	27.679611,	2.316829		222.974306,	3.396374		417.789391,	3.710582
	30.878440,	2.378714		226.116977,	3.403383		420.931295,	3.714329
	34.067233,	2.433228		229.259619,	3.410294		424.073196,	3.718048
	37.248411,	2.481964		232.402232,	3.417111		427.215091,	3.721740
	40.423677,	2.526053		235.544818,	3.423836		430.356983,	3.725405
	43.594260,	2.566325		238.687379,	3.430472		433.498870,	3.729043
	46.761071,	2.603408		241.829914,	3.437021		436.640753,	3.732655
	49.924800,	2.637782		244.972426,	3.443484		439.782631,	3.736241
	53.085978,	2.669829		248.114914,	3.449865		442.924506,	3.739801
	56.245022,	2.699852		251.257380,	3.456165		446.066377,	3.743336
	59.402261,	2.728099		254.399825,	3.462387		449.208244,	3.746847
	62.557962,	2.754775		257.542249,	3.468532		452.350107,	3.750333
	65.712341,	2.780049		260.684654,	3.474603		455.491966,	3.753794
	68.865575,	2.804066		263.827039,	3.480600		458.633822,	3.757232
	72.017811,	2.826947		266.969405,	3.486526		461.775674,	3.760646
	75.169172,	2.848797		270.111754,	3.492383		464.917523,	3.764038
	78.319763,	2.869708		273.254086,	3.498171		468.059369,	3.767406
	81.469670,	2.889759		276.396400,	3.503894		471.201211,	3.770752
	84.618968,	2.909018		279.538699,	3.509551		474.343049,	3.774075
	87.767723,	2.927548		282.680982,	3.515145	150th	477.484885,	3.777377
	90.915989,	2.945402		285.823250,	3.520677		480.626717,	3.780657
	94.063816,	2.962630	90th	288.965504,	3.526149		483.768546,	3.783916
	97.211245,	2.979273		292.107743,	3.531561		486.910373,	3.787153
30th	100.358313,	2.995371		295.249969,	3.536915		490.052196,	3.790370
	103.505052,	3.010960		298.392182,	3.542212		493.194016,	3.793566
	106.651492,	3.026070		301.534381,	3.547453		496.335834,	3.796741
	109.797658,	3.040731		304.676568,	3.552641		499.477648,	3.799897
	112.943572,	3.054968		307.818744,	3.557774		502.619460,	3.803033
	116.089255,	3.068807		310.960907,	3.562856		505.761270,	3.806149
	119.234724,	3.082269		314.103059,	3.567886		508.903076,	3.809246
	122.379997,	3.095374		317.245201,	3.572866		512.044880,	3.812324
	125.525087,	3.108141		320.387331,	3.577797		515.186681,	3.815383
	128.670009,	3.120587		323.529451,	3.582680		518.328480,	3.818424
	131.814773,	3.132727		326.671561,	3.587516		521.470277,	3.821446
	134.959391,	3.144578		329.813662,	3.592305		524.612071,	3.824450
	138.103873,	3.156151		332.955753,	3.597048		527.753862,	3.827436
	141.248227,	3.167461		336.097834,	3.601747		530.895652,	3.830404
	144.392463,	3.178519		339.239907,	3.606403		534.037439,	3.833355
	147.536587,	3.189336		342.381971,	3.611015		537.179223,	3.836288
	150.680606,	3.199922		345.524026,	3.615585		540.321006,	3.839204
	153.824528,	3.210287		348.666073,	3.620113		543.462786,	3.842104
	156.968357,	3.220440		351.808112,	3.624601		546.604564,	3.844986
	160.112098,	3.230390		354.950143,	3.629049		549.746340,	3.847852
	163.255758,	3.240144		358.092167,	3.633458		552.888114,	3.850702
	166.399341,	3.249711		361.234183,	3.637828		556.029886,	3.853536
	169.542850,	3.259097		364.376192,	3.642160		559.171656,	3.856353
	172.686291,	3.268310		367.518193,	3.646455		562.313424,	3.859155
	175.829666,	3.277355		370.660188,	3.650713		565.455190,	3.861941
	178.972979,	3.286238		373.802176,	3.654936		568.596955,	3.864712
	182.116233,	3.294966,		376.944158,	3.659123	180th	571.738717,	3.867468
	185.259431,	3.303543		380.086133,	3.663275		574.880477,	3.870208
	188.402576,	3.311975	120th	383.228102,	3.667393		578.022236,	3.872934
	191.545670,	3.320267		386.370065,	3.671477		581.163993,	3.875644
60th	194.688717,	3.328422		389.512022,	3.675528		584.305748,	3.878340
	197.831718,	3.336447		392.653974,	3.679547		587.447501,	3.881022
	200.974675,	3.344344		395.795919,	3.683533			

where $\mu(k, a_0)$ is given by (3.8). *In particular, when $a=1$ and $0<\alpha<1$, we have*

$$\frac{d}{da_0}\mu(k, a_0)>0 \qquad (0<\alpha<1) \tag{4.9}$$

so the rates of damping are decreasing.

Proof. a_0 is the radius of the caustic, $0<a_0<1$. As m increases in (3.8), a_0 decreases monotonically to zero [7]. Therefore, (4.7) and (4.8) hold. To see (4.9), we note that from (3.8),

$$\mu(k, a_0)=-\left[1+\frac{1}{3}\alpha(1-a_0^2)+\frac{1}{5}\alpha^2(1-a_0^2)^2+\cdots+\frac{1}{2n+1}\alpha^n(1-a_0^2)^n+\cdots\right].$$

So $\mu(k, a_0)$ decreases as a_0 decreases on the interval $(0, 1)$. □

So far in this section we have only dealt with the damping rates of eigenmodes. How good is formula (3.7) in approximating the wavenumber k? In [7], Keller and Rubinow did some calculations (using $\alpha=0$ and $\alpha=\infty$ in (3.6)) and noted that the numerical solutions are extremely close to the tabulated values of zeros of Bessel functions and their derivatives. Here we list two sets of values of k computed from (3.7), using

$$l=4,\ \alpha=0.5, \quad \text{for Table 5,}$$

$$l=4,\ \alpha=0.6+i\cdot 0.3, \quad \text{for Table 6.}$$

The reader can compare those values against, columns 3 and 5 in Table 3, respectively, and find that they are rather agreeable, the more so when the wavenumber becomes large. Here our fractional errors are considerably larger than those reported in [7] (for the energy conserving cases) at all frequencies.

Remark 4.1. For each m, the smallest (positive) solution k of the transcendental equation (3.6) (corresponding to $m=0$ or 1) is known to have the strongest "whispering gallery" property. Such a whispering gallery mode has the largest radius a_0 for its

TABLE 5
The wavenumber k as computed by (3.7) *for $l=4$, $\alpha=0.5$.*

m		m	
1	7.553060	6	24.013419
2	11.048664	7	27.194222
3	14.361846	8	30.366699
4	17.607830	9	33.533268
5	20.820315	10	36.695487

TABLE 6
The wavenumber k as computed by (3.7) *for $l=4$, $\alpha=0.6+i\cdot 0.3$.*

m		m	
1	7.180463	6	23.602581
2	10.652558	7	26.782572
3	13.957908	8	29.954486
4	17.200239	9	33.120648
5	20.410711	10	36.282563

caustic. Since a whispering gallery mode is "essentially supported" only on an annulus with outer radius a and inner radius a_0, as shown in Fig. 1. Therefore, a whispering gallery wave propagates right adjacent to $\partial\Omega$. Since there is boundary damping present at $\partial\Omega$, *will the whispering gallery mode get damped with the largest damping rate* (in the same family of eigenmodes, i.e., given fixed l)? The answer is *negative* as evidenced by the damping rates data given in column 2, Table 3, for $l=4$, with $\alpha=0.5$. The reason is that the wave "slides" along the boundary; thus the energy loss from damping, being "the friction in the perpendicular direction," is not as large as expected.

Remark 4.2. By adapting the work in [7, pp. 35-39] to our problem, it is not difficult to show that the eigenmode $\phi(x)$ has the asymptotic representation

$$(4.10)\qquad \phi(x)=\phi(r,\theta)=\begin{cases} [(kr)^2-l^2]^{-1/4}\cos\left\{[(kr)^2-l^2]^{1/2}-l\cos^{-1}\dfrac{l}{kr}\right. \\ \qquad\left.-\left[\dfrac{3\pi}{4}+\dfrac{1}{2}\arg\left(\dfrac{ka-\alpha\sqrt{(ka)^2-l^2}}{ka+\alpha\sqrt{(ka)^2-l^2}}\right)\right]\right\} \\ \qquad\qquad\text{if } r\geqq a_0, \\ \dfrac{1}{2}[l^2-(kr)^2]^{-1/4}\exp\left\{il\theta-l\cos h^{-1}\left(\dfrac{l}{kr}\right)+[l^2-(kr)^2]\right\}, \\ \qquad\qquad\text{if } r\leqq a_0. \end{cases}$$

The second representation can be regarded as the analytic continuation of the first one. This function is easily seen to be exponentially small inside the caustic (i.e., $r\leqq a_0$) when l becomes large.

Remark 4.3. Let the damping parameter α be positive and small. It is known that the eigenmodes (4.1), where λ is a solution of (4.2) and λ is close to λ_{lm} (cf. (4.3)), do *not* constitute a *complete* set of eigenfunctions of $L^2(\Omega)$. Another family of eigenmodes must be added to make a complete set. This family is given by

$$I_l(\lambda r)\, e^{\pm il\theta},$$

where I_l is the modified Bessel function of order l, and λ is a (unique) real positive root of

$$I_l(\lambda)-\alpha I_l'(\lambda)=0.$$

Such an eigenmode decays exponentially in time *without oscillations*, i.e.,

$$w(x,t)=e^{-\lambda t}I_l(\lambda r)\,e^{\pm il\theta},\qquad \lambda>0,$$

satisfies (1.1) and (1.2). We will call such modes *overdamped* modes. Our wave method developed here *does not* cover such overdamped modes. Such eigenmodes seem to be intimately connected with the complex or imaginary rays [6], [7] and with the second representation formula in (4.10). We hope to be able to treat them in a separate article.

5. The case of rectangular geometry. Quinn and Russell [11] first studied the problem (1.1), (1.2) on a rectangular domain, which initiated a series of papers by Chen and others. We describe their setup of the problem and results. Let Ω be a rectangle:

$$\Omega=\{(x_1,x_2)|0<x_1<a,0<x_2<b\}$$

throughout the rest of this section. On $\partial\Omega$, the boundary conditions are prescribed as

$$(5.1)\qquad \begin{aligned} &w=0 && \text{on } \mathscr{S}_1\colon 0\leqq x\leqq a,\quad y=0,\quad t\geqq 0,\\ &\frac{\partial w}{\partial t}+\alpha\frac{\partial w}{\partial x}=0 && \text{on } \mathscr{S}_2\colon 0\leqq y\leqq b,\quad x=a,\quad t\geqq 0,\quad \Re\alpha\geqq 0,\\ &\frac{\partial w}{\partial t}+\beta\frac{\partial w}{\partial y}=0 && \text{on } \mathscr{S}_3\colon 0\leqq x\leqq a,\quad y=b,\quad t\geqq 0,\quad \Re\beta\geqq 0,\\ &w=0 && \text{on } \mathscr{S}_4\colon 0\leqq y\leqq b,\quad x=0,\quad t\geqq 0.\end{aligned}$$

In [11], Quinn and Russell let $a=b=\pi$, and α, β real and small. They showed that under such circumstances there exists a two-parameter family of eigenmodes:

$$w(x,t)=e^{-i\lambda t}\phi(x),$$

$$(5.2)\qquad \lambda=\sqrt{l^2+m^2}-\frac{i}{\pi}\left(\frac{\alpha l^2+\beta m^2}{l^2+m^2}\right)+O(\alpha^2+\beta^2),\qquad l,m=1,2,3,\cdots$$

and another one-parameter family of overdamped eigenmodes:

$$(5.3)\qquad \begin{gathered} w(x,t)=e^{-\lambda_j t}\phi(x),\\ 0<\lambda_1<\lambda_2<\cdots<\lambda_j<\cdots\to+\infty.\end{gathered}$$

We will leave out the study of overdamped modes in the present paper.

In [11], Quinn and Russell also gave explicit forms for the $\phi(x)$ in (5.3) in their paper. (They had an extra factor $\frac{1}{2}$ in their expression immediately before their Lemma 6.2, which has been corrected by us in (5.2).)

We now study the eigenfrequencies for the rectangular geometry and boundary conditions (5.1) by the wave method.

For the rectangle, Keller and Rubinow [7, pp. 57–59] pointed out that there are four normal congruences of waves, i.e., $N=4$, and the covering space is again topologically a torus. The four normal congruences of waves are depicted in Figs. 6–9.

Consider the family of rays in Fig. 6. The geometrical optics expansion in § 2 reduces to

$$(5.4)\qquad \phi(x)=A_1\exp\left(ik(\eta_1x_1+\eta_2x_1)\right),$$

where

$$(5.5)\qquad \begin{gathered}\eta_1=\cos\theta,\qquad \eta_2=\sin\theta\\ (\theta=\text{the angle formed between the ray and the } x_1 \text{ axis}).\end{gathered}$$

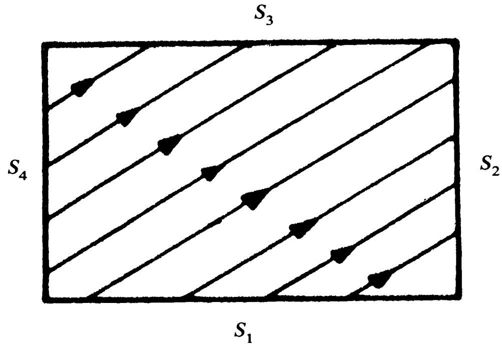


FIG. 6. *Rays of normal congruence* I. (*Reprinted from* [7], *with permission.*)

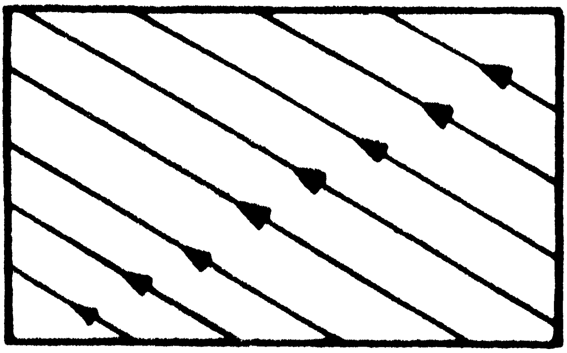


FIG. 7. *Rays of normal congruence* II. (*Reprinted from* [7], *with permission.*)

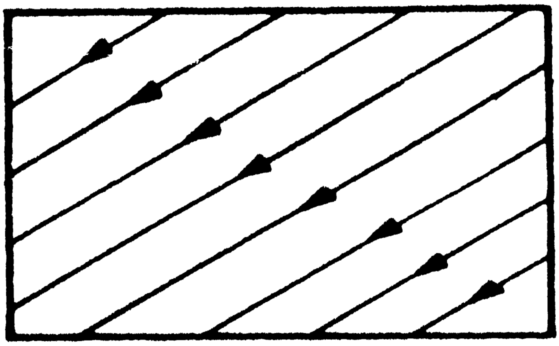


FIG. 8. *Rays of normal congruence* III. (*Reprinted from* [7], *with permission.*)

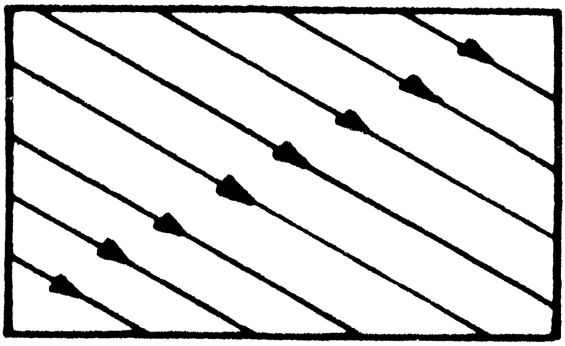


FIG. 9. *Rays of normal congruence* IV. (*Reprinted from* [7], *with permission.*)

Upon hitting side $\mathcal{S}_2$, reflection happens. Utilizing (2.17), we obtain the reflected wave of normal congruence II

$$\left[\frac{(-1+\alpha\eta_1)}{(1+\alpha\eta_1)}e^{ik\eta_1 a}A_1\right]\exp\left(ik[-\eta_1(x_1-a)+\eta_2x_2]\right), \tag{5.6}$$

the one as shown in Fig. 7. This wave moves left, hits side $\mathcal{S}_4$, makes a reflection there, and returns to the same direction as in Fig. 6. The reflection of wave (5.6), according to the last boundary condition in (5.1), gives rise to a wave of normal congruence I:

$$-\left[\frac{(-1+\alpha\eta_1)}{(1+\alpha\eta_1)}e^{2ik\eta_1 a}A_1\right]\exp(ik(\eta_1x_1+\eta_2x_2)). \tag{5.7}$$

In order for an eigenmode to form, resonance must happen so the two waves (5.4) and (5.7) must be in phase

$$\arg\left[-\frac{(-1+\alpha\eta_1)}{(1+\alpha\eta_1)}e^{2ik\eta_1 a}\right]=2l\pi, \qquad l\in\mathbb{Z},$$

$$\arg\left(\frac{1-\alpha\eta_1}{1+\alpha\eta_1}\right)+2k\eta_1 a=2l\pi, \qquad l\in\mathbb{Z}. \tag{5.8}$$

This is equivalent to applying (2.22) to a first closed path, $j=1$, on the torus, the covering space of wave motion.

Similarly, if we track the motions of normal congruences and III and II of rays as shown in Figs. 7 and 8, beginning from

$$\phi(x)=A_2\exp(ik[-\eta_1x_1-\eta_2(x_2-b)])$$

in Fig. 8, we see that after first reflection at $\mathcal{S}_1$ and second reflection at $\mathcal{S}_3$, this wave becomes

$$\left[\frac{(1-\beta\eta_2)}{(1+\beta\eta_2)}e^{2ik\eta_2 b}A_2\right]\exp(ik[-\eta_1x_1-\eta_2(x_2-b)]).$$

So, again, we deduce that

$$\arg\left(\frac{1-\beta\eta_2}{1+\beta\eta_2}\right)+2k\eta_2 b=2m\pi, \qquad m\in\mathbb{Z}^+. \tag{5.9}$$

This is equivalent to applying (2.22) to the second independent closed path, $j=2$, on the torus. From (5.5),

$$\eta_1^2+\eta_2^2=1; \tag{5.10}$$

therefore, the wavenumber k satisfies

$$k^2=\frac{1}{4a^2}\left[2l\pi-\arg\left(\frac{1-\alpha\eta_1}{1+\alpha\eta_1}\right)\right]^2+\frac{1}{4b^2}\left[2m\pi-\arg\left(\frac{1-\beta\eta_2}{1+\beta\eta_2}\right)\right]^2, \qquad l,m\in\mathbb{Z}, \tag{5.11}$$

subject to

$$\eta_1^2+\eta_2^2=1.$$

Note that once l and m are given, k is determinable from (5.8)-(5.10). So the solutions k form a two-parameter family.

A special case is where $\alpha=\beta=0$, then

$$\arg\frac{1-\alpha\eta_1}{1+\alpha\eta_1}=0, \qquad \arg\frac{1-\beta\eta_2}{1+\beta\eta_2}=0,$$

and we have

$$k^2=\left(\frac{l^2}{a^2}+\frac{m^2}{b^2}\right)\pi^2, \qquad l,m\in\mathbb{Z}. \tag{5.12}$$

This agrees entirely with the separation of variables approach because the boundary conditions (5.1) reduce to the Dirichlet condition, which is known to have eigenfrequencies k satisfying (5.12).

Similarly, if $\alpha=\beta=+\infty$, then

$$\arg\frac{1-\alpha\eta_1}{1+\alpha\eta_1}=\pm\pi, \qquad \arg\frac{1-\beta\eta_2}{1+\beta\eta_2}=\pm\pi$$

and (5.11) yields

$$k^2=\left[\frac{1}{4a^2}(2l+1)^2+\frac{1}{4b^2}(2m+1)^2\right]\pi^2, \qquad l, m\in\mathbb{Z}.$$

This determines the eigenfrequencies k for a mixed Dirichlet-Neumann boundary value problem which are also well understood.

Now let us compute the damping rates of eigenmodes. We note that for normal congruence I and II waves, *a complete cycle of wave motion* takes $2a/\eta_1$ units of time because $2a/\eta_1$ is the distance traveled by these waves. The total amount of damping that the waves have endured is

$$\ln\left|\frac{1-\alpha\eta_1}{1+\alpha\eta_1}\right|.$$

Therefore, the rate of damping per unit time for such a cycle of wave motion is

$$\frac{1}{(2a/\eta_1)}\ln\left|\frac{1-\alpha\eta_1}{1+\alpha\eta_1}\right|.$$

Similarly, the rate of damping per unit time for a cycle of wave motion as characterized by (5.9) is

$$\frac{1}{(2b/\eta_2)}\ln\left|\frac{1-\beta\eta_2}{1+\beta\eta_2}\right|.$$

Hence, the rate of damping per unit time for a wave to go through the complete sequence of motion

$$\text{waves (I, II)}\rightarrow\text{waves (II, III)}\rightarrow\text{waves (I, II)}$$

is

$$\left[\frac{1}{(2a/\eta_1)}\ln\left|\frac{1-\alpha\eta_1}{1+\alpha\eta_1}\right|+\frac{1}{(2b/\eta_2)}\ln\left|\frac{1-\beta\eta_2}{1+\beta\eta_2}\right|\right]. \tag{5.13}$$

This expression is obtained based upon the understanding that vertical and horizontal wave motions are independent of each other. Because of the amplitudes' multiplicative effect, the total rate of damping (as the logarithm of gain ratio) must be additive of damping rates in the separate horizontal and vertical directions. Also note that on a rectangle, there does not exist any caustics, so (5.13) will suffice to provide the damping rate per unit time, from the same ideas as in the circular geometry.

Any other complete sequences of wave motions (such as combinations of (III, II), (I, IV)) will also give the same damping rate as (5.13).

COROLLARY 3. *When α and β are real nonnegative and small, $a=b=\pi$, the damping rate per unit time* (5.13) *is*

$$-\frac{1}{\pi}\left(\frac{\alpha l^2+\beta m^2}{l^2+m^2}\right)+O(\alpha^2+\beta^2), \qquad l, m\in\mathbb{Z}. \tag{5.14}$$

Proof. When α and β are real and small

$$1-\alpha\eta_1>0, \qquad 1-\beta\eta_2>0$$

because $|\eta_1|\leqq 1$, $|\eta_2|\leqq 1$. Thus

$$\arg\left(\frac{1-\alpha\eta_1}{1+\alpha\eta_1}\right)=0, \qquad \arg\left(\frac{1-\beta\eta_2}{1+\beta\eta_2}\right)=0.$$

Therefore, from (5.8) and (5.9), we obtain

$$\eta_1=\frac{l}{k}, \qquad \eta_2=\frac{m}{k}. \tag{5.15}$$

Substituting (5.15) into (5.13), for small α and β, we have

$$\begin{aligned}(5.13)&=\frac{\eta_1}{2a}[-2\alpha\eta_1+O(\alpha^2\eta_1^2)]+\frac{\eta_2}{2b}[-2\beta\eta_2+O(\beta^2\eta_2^2)]\\ &=-\frac{1}{\pi}\{\alpha\eta_1^2+\beta\eta_2^2+O(\alpha^2+\beta^2)\}\\ &=-\frac{1}{\pi}\left(\frac{\alpha l^2+\beta m^2}{l^2+m^2}\right)+O(\alpha^2+\beta^2).\end{aligned}$$ □

Therefore, we see that the damping rate (5.14) agrees with (the dominant) imaginary part of λ in (5.2) given by Quinn and Russell, for small α and β. But our formulas (5.13) and (5.11) (subject to (5.8)-(5.10)) are more general because they apply to almost all range of values of α, β, $\Re\alpha\geqq 0$, $\Re\beta\geqq 0$.

There have been many physical arguments used by Keller and Rubinow in [7] and by us in this paper. These arguments are all based on the accepted standards of wave propagation and optics and have been rigorously established in [10] by Maslov and coworkers; therefore, they yield asymptotically accurate solutions.

REFERENCES

[1] G. CHEN, *Energy decay estimates and control theory for the wave equation in a bounded domain*, Ph.D. thesis, University of Wisconsin, Madison, WI, May 1977.

[2] ———, *Energy decay estimates and exact boundary value controllability for the wave equation in a bounded domain*, J. Math. Pures Appl., 58 (1979), pp. 249-273.

[3] G. CHEN, M. P. COLEMAN, AND H. H. WEST, *Pointwise stabilization in the middle of the span for second-order systems, nonuniform and uniform decay results*, SIAM J. Appl. Math., 47 (1987), pp. 751-780.

[4] G. CHEN AND J. ZHOU, *The wave propagation method for the analysis of boundary stabilization in vibrating structures*, SIAM J. Appl. Math., 50 (1990), pp. 1254-1283.

[5] J. B. KELLER, R. M. LEWIS, AND B. D. SECKLER, *Asymptotic solution of some diffraction problems*, Comm. Pure Appl. Math., 9 (1956), pp. 207-265.

[6] J. B. KELLER, *A geometrical theory of diffraction*, in Proc. Symposia in Applied Mathematics, Vol. VII, McGraw-Hill, New York, 1958, pp. 27-52.

[7] J. B. KELLER AND S. I. RUBINOW, *Asymptotic solution of eigenvalue problems*, Ann. Phys., 9 (1960), pp. 24-75.

[8] J. E. LAGNESE, *Decay of solutions of wave equations in a bounded region with boundary dissipation*, J. Differential Equations, 50 (1983), pp. 163-182.

[9] A. MAJDA, *Disappearing solutions for the dissipative wave equation*, Indiana Univ. Math. J., 24 (1975), pp. 1119-1133.

[10] V. P. MASLOV AND M. V. FEDORIUK, *Semi-classical Approximation in Quantum Mechanics*, Reidel, Boston, MA, 1981.

[11] J. P. QUINN AND D. L. RUSSELL, *Asymptotic stability and energy decay rates for solutions of hyperbolic equations with boundary damping*, Proc. Roy. Soc. Edinburgh Sect. A, 77 (1977), pp. 97-127.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 3, pp. 678-701, May 1991

010

NONSMOOTH PROBLEMS WITH CONFLICTING CONTROLS*

J. WARGA†

Abstract. This paper considers nonsmooth problems defined by ordinary differential equations and involving conflicting controls of two players. These problems differ from differential games in that the second player is informed of the first player's choice of control function before making his own choice. It attempts, with only partial success, to generalize the results obtained for C^1 problems in [J. Warga, *Optimal Control of Differential and Functional Equations*, Chaps. IX and X, Academic Press, 1972]. The results are in the form of a familiar alternative: either the problem is "strongly controllable" at a first player's relaxed control σ_0 or certain maximum principles and transversality conditions are valid. Two models are discussed in which the controls of the second player are relaxed controls, respectively, hyperrelaxed controls.

Key words. differential equations, conflicting controls, relaxed controls, hyperrelaxed controls, necessary conditions, derivate containers

AMS(MOS) subject classifications. 49B10, 49B40, 49E15

1. Introduction. We will consider problems with conflicting controls that bear some formal resemblance to differential games but provide the second player with a priori information about the first player's decisions. A special case of these problems are minimax problems of the form $\inf_u \sup_v$, where u and v are the controls of the first and second players. Specifically, consider functions y, $\tilde{y}$, and $\hat{y}=(y,\tilde{y})$ satisfying the differential equations

$$(1.1)\qquad \begin{aligned} y(t)&=\int_{t_0}^{t} f(s,y(s),u(s))\,ds,\\ \tilde{y}(t)&=\int_{t_0}^{t} \tilde{f}(s,\hat{y}(s),u(s),v(s))\,ds \quad \forall t\in T:=[t_0,t_1], \end{aligned}$$

in which R and R_P are compact metric spaces and $u:T\to R$ and $v:T\to R_P$ are (Lebesgue) measurable functions or, more generally, measurable selections of measurable set-valued mappings $t\to R^{\#}(t)$ and $t\to R_P^{\#}(t)$. For given functions

$$h=(h^1,\cdots,h^m):\mathbb{R}^n\to\mathbb{R}^m,\qquad \hat{h}:\mathbb{R}^n\times\mathbb{R}^p\to\mathbb{R}$$

(where $\mathbb{R}$ denotes the space of real numbers), we set

$$\varphi(u)=(\varphi^1,\cdots,\varphi^m)(u)=h(y(t_1)),\qquad \Phi(u)(v)=\hat{h}(\hat{y}(t_1))$$

and consider the restricted attainable set

$$(1.2)\qquad \mathcal{F}=\{\varphi(u)\mid\Phi(u)(v)\leqq 0\ \forall v\}.$$

If u_0 yields the minimum of $\varphi^1(u)$ subject to $\varphi^j(u)=0$ for $j=2,\cdots,m$ and $\Phi(u)(v)\leqq 0$ for all v, then clearly $\varphi(u_0)$ is on the boundary of $\mathcal{F}$. This cannot be the case if φ is either *controllable* at u_0, i.e., some neighborhood of $\varphi(u_0)$ is contained in $\mathcal{F}$, or a fortiori, if φ is *strongly controllable* at u_0, i.e., there exists $\beta>0$ such that some neighborhood of $\varphi(u_0)$ is contained in

$$(1.3)\qquad \mathcal{F}_\beta:=\{\varphi(u)\mid\Phi(u)(v)\leqq-\beta\ \forall v\}.$$

* Received by the editors October 2, 1989; accepted for publication (in revised form) January 31, 1990. This research was partially supported by National Science Foundation grant DMS-8619002.

† Department of Mathematics, Northeastern University, Boston, Massachusetts 02115.

(We might mention, in passing, that the restriction $\hat{h}(\hat{y}(t_1)) \leqq 0$ is more general than it may appear. Thus, e.g., a restriction of the form $h^{\#}(\hat{y}(t_1)) \in A \subset \mathbb{R}^a$, with A a convex body, is equivalent to $\hat{h}(\hat{y}(t_1)) \leqq 0$ if we set $h = \phi \circ h^{\#}$, where ϕ is a gauge function of A. For a more detailed discussion, see [6, p. 29].)

A similar problem was studied in [4, Chaps. IX and X] under the assumption that the functions $f(t, \cdot, r)$, $\tilde{f}(t, \cdot, r, r_P)$, h and $\hat{h}$ are C^1. It was shown there—and the proof remains valid under the weaker assumptions of this paper—that there are two ways of compactifying the problem, corresponding to different "physical" models. The first of these ways replaces the conflicting ordinary controls u and v with relaxed controls σ and π so that (1.1) is replaced by

$$\text{(1.4)}\qquad \begin{aligned} \hat{y}(t) &= \int_{t_0}^{t} \hat{f}(s, \hat{y}(s), \sigma(s), \pi(s))\, ds \\ &= \int_{t_0}^{t} ds \int \hat{f}(s, \hat{y}(s), r, r_P)\sigma(s)(dr) \times \pi(s)(dr_P) \quad \forall t \in T, \end{aligned}$$

where $\hat{f} = (f, \tilde{f})$. This properly models two kinds of problems: problems with additively coupled (or separated) controls in which

$$\hat{f}(t, \hat{\mathbf{y}}, r, r_P) = \hat{f}_1(t, \hat{\mathbf{y}}, r) + \hat{f}_2(t, \hat{\mathbf{y}}, r_P)$$

(where the boldfaced letter $\hat{\mathbf{y}}$ represents a point in the space $\mathbb{R}^{n+p}$ to distinguish it from the function $\hat{y}$ with values in $\mathbb{R}^{n+p}$); and problems in which the adversary (who controls v) does not have perfect means of detecting the exact value of $u(t)$ but only detects the distribution of values of $u(\cdot)$ over short intervals of time. The second kind of compactification replaces $u(\cdot)$ with the corresponding relaxed controls $\sigma(\cdot)$ while replacing $v(\cdot)$ with the corresponding *hyperrelaxed* controls $\pi: T \times R \to \operatorname{rpm}(R_P)$, where $\operatorname{rpm}(R_P)$ represents the set of Radon probability measures on R_P. This latter type of compactification properly models problems with nonseparated controls and with perfect observation capabilities of the v-controller, and replaces the sets $\mathcal{F}$ and $\mathcal{F}_\beta$ of (1.2) and (1.3) with their closures.

In the present paper we primarily study, in Theorem 2.4, the controllability of the problem when it is subjected to the first kind of compactification involving relaxed controls of both opponents. However, we also derive a weaker result, Theorem 2.5, pertaining to problems with hyperrelaxed adverse controls. In both cases, we assume that $f(t, \cdot, r)$, $\hat{f}(t, \cdot, r, r_P)$, h, and $\hat{h}$ are Lipschitz continuous but not necessarily C^1. We also require the additional "semi-C^1" assumption (2.3.5)

$$\hat{f}(t, \hat{\mathbf{y}}, r, r_P) = \hat{f}_1(t, \hat{\mathbf{y}}, r) + \hat{f}_2(t, \hat{\mathbf{y}}, r, r_P),$$

where $\hat{f}_2(t, \cdot, r, r_P)$ is assumed to be C^1. Our results are in the form of an alternative: either the problem is strongly controllable or certain maximum principles and transversality conditions are valid. The maximum principles (or, more properly, minimum principles) appear in different forms. In Theorem 2.4, the "integral" form 2.4(b2.1) of the maximum principle is analogous to those obtained for many other nonsmooth problems. On the other hand, our first "pointwise" principle 2.4(b2.2) involves an integral of a function with values in L^2 provided with a "weak" norm, while each version of the second one, 2.4(b2.3), requires a special assumption (verifiable a priori). The maximum principle 2.4(b2.2) involves certain considerations about an interchange of the order of integration that bear a superficial resemblance to some aspects of stochastic integration [3, Thm. 3A, p. 79].

We define the framework of our problems and the corresponding results, Theorems 2.4 and 2.5, in § 2. In that section we also include an abstract result, Theorem 2.6, which we use in proving Theorems 2.4 and 2.5 and which we expect to have other applications in optimal control. In § 3 we comment on the maximum principles of Theorems 2.4 and 2.5 and on the "semi-C^1" assumption (2.3.5). The proofs appear in § 4.

2. Definitions and results.

2.1. Preliminary definitions and notation. We denote by $\bar{A}$ or cl A, A°, ∂A, co A, and $\overline{\text{co}}\, A$ the closure, interior, boundary, convex hull, and convex closure of A. We endow the spaces $\mathbb{R}^a$ with the Euclidean norm and the spaces $\mathscr{L}(\mathbb{R}^a, \mathbb{R}^b)$ (of $b \times a$ matrices) with the (strictly convex) norm $|(M_{ij})| := [\sum_{i,j} M_{ij}^2]^{1/2}$. We represent the elements of any $\mathbb{R}^a$ as column vectors and denote their transpose by the superscript T. We denote by $\bar{B}_{\mathscr{Y}}$ the closed unit ball in a normed vector space $\mathscr{Y}$ and by $\bar{B}_m$ the closed unit ball in $\mathbb{R}^m$. If ν is a Radon measure on a compact metric space X, $x \in X$, and $g: X \to \mathbb{R}^a$ continuous, we write supp(ν) for the support of ν, δ_x for the Dirac measure at x, and $g(\nu)$ for $\int g(x)\nu(dx)$.

We use some of the notation, definitions, and results of [4, Chaps. IX and X], which we summarize in the present context. Let $T := [t_0, t_1] \subset \mathbb{R}$, μ be the Lebesgue measure on T; R and R_P compact metric spaces, $\hat{R} := R \times R_P$; and $R^\#: T \to 2^R$ and $R_P^\#: T \to 2^{R_P}$ μ-measurable mappings whose values are nonempty closed sets. We write $C(T, \mathbb{R}^a)$ $[AC(T, \mathbb{R}^a)]$ for the spaces of all continuous (absolutely continuous) functions from T to $\mathbb{R}^a$ with the sup norm, and $L^q(\mu, \mathscr{X})$ for the Banach space of μ-measurable functions with values in a Banach space $\mathscr{X}$ and with the usual norm $|\cdot|_q < \infty$. We denote by frm (X) the vector space of all Radon measures on X, and by frm$^+$ (X) and rpm (X) its subsets consisting of nonnegative and probability measures, each with the weak* topology of $C(X)^*$ (the topological dual of $C(X) = C(X, \mathbb{R})$).

We denote by $\mathscr{S}^\#$ the collection of *relaxed controls*, i.e., μ-measurable mappings $\sigma: T \to \text{rpm}(R)$ with $\sigma(t)(R^\#(t)) = 1$ μ-almost everywhere and with the weak* topology of $L^1(\mu, C(R))^*$. We similarly define $\mathscr{S}_P^\#$, respectively $\hat{\mathscr{S}}^\#$, with R, $R^\#(t)$ replaced by R_P, $R_P^\#(t)$, respectively, $\hat{R}$, $R^\#(t) \times R_P^\#(t)$. The sets $\mathscr{S}^\#$, $\mathscr{S}_P^\#$, and $\hat{\mathscr{S}}^\#$ are compact and convex. We write $\mathscr{P}^\#$ for the collection of *hyperrelaxed controls*, i.e., Borel measurable mappings $\pi: T \times R \to \text{rpm}(R_P)$ with $\pi(t, r)(R_P^\#(t)) = 1$ for all $r \in R$ μ-almost everywhere. For each $\sigma \in \mathscr{S}^\#$ and $\pi \in \mathscr{P}^\#$, we write $\sigma \otimes \pi$ for the unique element of $\hat{\mathscr{S}}^\#$ such that

$$\int_{t_0}^{t_1} dt \int \varphi(t, r, r_P)\sigma \otimes \pi(t)(d(r, r_P)) = \int_{t_0}^{t_1} dt \int \sigma(t)(dr) \int \varphi(t, r, r_P)\pi(t, r)(dr_P)$$
$$\forall \varphi \in L^1(\mu, C(\hat{R})).$$

There appears to be no useful way of defining a compact metric topology for $\mathscr{P}^\#$ so as to render the mapping $\pi \to \sigma \otimes \pi : \mathscr{P}^\# \to \hat{\mathscr{S}}^\#$ continuous for every $\sigma \in \mathscr{S}^\#$. However, if we choose any denumerable subset $\mathscr{S}'$ of $\mathscr{S}^\#$, then we can define [4, Def. X.2.1]

(a) a set $\tilde{\mathscr{P}}$ of equivalence classes in $\mathscr{P}^\#$ such that

$$\sigma \otimes \pi_1(t) = \sigma \otimes \pi_2(t) \quad \mu\text{-a.e.} \quad \forall \sigma \in \text{co}\, \mathscr{S}'$$

if π_1 and π_2 belong to the same equivalence class, and

(b) a compact metric topology for $\tilde{\mathscr{P}}$ such that the mapping $\pi \to \sigma \otimes \pi : \tilde{\mathscr{P}} \to \hat{\mathscr{S}}^\#$ is continuous for every $\sigma \in \text{co}\, \mathscr{S}'$.

We will use the symbol $\mathscr{P}$ to denote either $\mathscr{S}_P^\#$ or $\tilde{\mathscr{P}}$ and, for $\sigma \in \mathscr{S}^\#$ and $\pi \in \mathscr{S}_P^\#$, we will write $\sigma \otimes \pi(t)$ for $\sigma(t) \times \pi(t)$. We observe that, with these definitions, the mapping $\pi \to \sigma \otimes \pi : \mathscr{P} \to \hat{\mathscr{S}}^{\#}$ is continuous for every choice of $\sigma \in \operatorname{co} \mathscr{S}'$, where $\mathscr{S}'$ is an arbitrary denumerable subset of $\mathscr{S}^{\#}$ and $\tilde{\mathscr{P}}$ is defined accordingly. If $g : R \times R_P \to \mathbb{R}^a$ is continuous, we write $g(\sigma \otimes \pi(t))$ or $g(\sigma(t), \pi(t))$ for $\int \sigma(t)(dr) \int g(r, r_P)\pi(t)(dr_P)$ if $\pi \in \mathscr{S}_P^{\#}$ and $g(\sigma \otimes \pi(t))$ for $\int \sigma(t)(dr) \int g(r, r_P)\pi(t, r)(dr_P)$ if $\pi \in \tilde{\mathscr{P}}$. We observe that for any continuous $g : R \to \mathbb{R}^a$ we have $g(\sigma \otimes \pi(t)) = g(\sigma(t))$ for all $\pi \in \mathscr{P}$. We also write $\int_w \varphi(\pi)\omega(d\pi)$ for the integral of a function $\varphi : \mathscr{P} \to (L^2(\mu, \mathbb{R}), |\cdot|_w)$, where $|\cdot|_w$ is a "weak" norm on $L^2(\mu, \mathbb{R})$ that can be defined as follows. Let $\{x_1, x_2, \cdots\}$ be a dense denumerable subset of $L^2(\mu, \mathbb{R})$ and

$$|\ell|_w := \sum_{i=1}^{\infty} \frac{2^{-i}}{1+|x_i|} |\langle \ell, x_i \rangle| \quad \forall \ell \in L^2(\mu, \mathbb{R}) = (L^2(\mu, \mathbb{R}))^*.$$

Then $|\cdot|_w$ is a norm on $L^2(\mu, \mathbb{R})$, and it induces the weak topology (=the weak* topology) on any bounded subset of $L^2(\mu, \mathbb{R}) = (L^2(\mu, \mathbb{R}))^*$ [4, Thm. I.3.11].

We will also use the concepts of a derivate container and of Clarke's generalized Jacobian [2, p. 70].

DEFINITION 2.2. Let $W \subset \mathbb{R}^a$ be open and $g : W \to \mathbb{R}^b$ locally Lipschitzian.

(2.2.1) We say that a collection $\{\Lambda^\varepsilon g(x)\} \mid \varepsilon > 0\}$, also denoted by $\Lambda^\varepsilon g(x)$, of closed bounded subsets of $\mathscr{L}(\mathbb{R}^a, \mathbb{R}^b)$ is a *derivate container* for g at x if

$$\Lambda^\varepsilon g(x) \subset \Lambda^{\varepsilon'} g(x) \quad \text{for } \varepsilon < \varepsilon'$$

and there exists a sequence $(g_i : W \to \mathbb{R}^b)$ of C^1 functions converging uniformly to g and such that

$$g_i'(y) \in \Lambda^\varepsilon g(x) \quad \text{if } y \in W, \quad |y - x| \leqq \varepsilon, \quad i \geqq 1/\varepsilon.$$

We set

$$\Lambda g(x) = \bigcap_{\varepsilon > 0} \Lambda^\varepsilon g(x)$$

for every derivate container $\Lambda^\varepsilon g(x)$, and refer to $\Lambda g(x)$, as well as to $\Lambda^\varepsilon g(x)$, as a derivate container. Since the sets

$$\Lambda_1^\varepsilon g(x) = \operatorname{cl}\{g_i'(y) \mid y \in W, |y - x| \leqq \varepsilon, i \geqq 1/\varepsilon\} \subset \Lambda^\varepsilon g(x)$$

also define a (better) derivate container, and the set-valued function $x \to \Lambda_1^\varepsilon g(x)$ is upper semicontinuous, we henceforth assume that the derivate containers that we use have this upper semicontinuity property.

(2.2.2) Let $D = \{y \in W \mid g'(y) \text{ exists}\}$,

$$\partial^\varepsilon g(x) = \overline{\operatorname{co}}\{g'(y) \mid y \in D, |y - x| \leqq \varepsilon\}, \quad \partial g(x) = \bigcap_{\varepsilon > 0} \partial^\varepsilon g(x).$$

Then $\partial g(x)$ coincides with Clarke's generalized Jacobian and, by [6, Thm. 2.5], $\partial^\varepsilon g$ is a derivate container that is upper semicontinuous.

Assumption 2.3. Let $V \subset \mathbb{R}^n$ and $\tilde{V} \subset \mathbb{R}^p$ be open, $\hat{V} = V \times \tilde{V}$, and the functions

$$f : T \times V \times R \to \mathbb{R}^n, \qquad \tilde{f} : T \times \hat{V} \times R \times R_P \to \mathbb{R}^p,$$

$$\hat{f} = (f, \tilde{f}) : T \times \hat{V} \times R \times R_P \to \mathbb{R}^{n+p}, \quad h : V \to \mathbb{R}^m, \quad \hat{h} : \hat{V} \to \mathbb{R}$$

such that

(2.3.1) $\hat{f}(\cdot, \hat{y}, r, r_P)$ are μ-measurable;

(2.3.2) $\hat{f}(t, \cdot, \cdot, \cdot)$ are continuous;

and there exist $c \geqq 0$ and a μ-integrable $\psi : T \to [0, \infty)$ such that

(2.3.3) $\hat{f}(t,\cdot,r,r_P)$ have $\psi(t)$, and h and $\hat{h}$ have c, as a common bound and a common Lipschitz constant;

(2.3.4) the differential equation

$$\hat{y}(t)=\int_{t_0}^{t}\hat{f}(s,\hat{y}(s),\sigma\otimes\pi(s))\,ds\quad\forall t\in T$$

has a unique solution $\hat{y}(\sigma\otimes\pi)(\cdot)=(y(\sigma),\tilde{y}(\sigma\otimes\pi))(\cdot)$ for every choice of $\sigma\in\mathcal{S}^{\#}$ and $\pi\in\mathcal{P}$;

(2.3.5) $\hat{f}(t,\hat{\mathrm{y}},r,r_P)=\hat{f}_1(t,\hat{\mathrm{y}},r)+\hat{f}_2(t,\hat{\mathrm{y}},r,r_P)$, where $\hat{f}_1$ and $\hat{f}_2$ satisfy the same conditions as $\hat{f}$ and, in addition, $\hat{f}_2$ has a partial derivative $\mathcal{D}_2\hat{f}_2$ with respect to the second argument of $\hat{f}_2$, and the function

$$(\hat{\mathrm{y}},r,r_P)\to\mathcal{D}_2\hat{f}_2(t,\hat{\mathrm{y}},r,r_P):\hat{V}\times R\times R_P\to\mathcal{L}(\mathbb{R}^{n+p},\mathbb{R}^{n+p})$$

is continuous for each $t\in T$.

We can now state our principal results. Theorem 2.4 deals with the case of adverse relaxed controls and Theorem 2.5 with the case of adverse hyperrelaxed controls. We write $\partial f(t,w,r)$ and $\partial\hat{f}(t,\hat{w},r,r_P)$ for the generalized Jacobians of $f(t,\cdot,r)$ at w and of $\hat{f}(t,\cdot,r,r_P)$ at $\hat{w}$.

THEOREM 2.4. *Let $\sigma_0\in\mathcal{S}^{\#}$ be such that $\hat{h}(\hat{y}(\sigma_0\otimes\pi)(t_1))\leqq 0$ for all $\pi\in\mathcal{S}^{\#}_P$, and let*

$$y_0(t):=y(\sigma_0)(t),\qquad \hat{y}_0(\pi)(t):=\hat{y}(\sigma_0\otimes\pi)(t).$$

Let Assumption 2.3 be satisfied and Λh and $\Lambda\hat{h}$ be derivate containers for h and $\hat{h}$. Then either

(a) *there exists $\beta>0$ such that*

$$h(y_0(t_1))+\beta\bar{B}_m\subset\{h(y(\sigma)(t_1))\,|\,\sigma\in\mathcal{S}^{\#},\ \hat{h}(\hat{y}(\sigma\otimes\pi)(t_1))\leqq-\beta\ \ \forall\pi\in\mathcal{S}^{\#}_P\}$$

or

(b) *there exist*

$$\ell^1\in\mathbb{R}^m,\quad \omega\in\mathrm{frm}^+(\mathcal{S}^{\#}_P),\quad z\in AC(T,\mathbb{R}^n),\quad \hat{z}:\mathcal{S}^{\#}_P\to AC(T,\mathbb{R}^{n+p})$$

such that

(b1)
$$0<|\ell^1|+\omega(\mathcal{S}^{\#}_P)\leqq 1$$

$$z(t)^T\in\ell^{1T}\Lambda h(y_0(t_1))+\int_t^{t_1}z(s)^T\,\partial f(s,y_0(s),\sigma_0(s))\,ds\quad\forall t\in T$$

$$\hat{z}(\pi)\in\overline{\mathrm{co}}\Big\{\zeta\in AC(T,\mathbb{R}^{n+p})\,|\,\zeta(t)^T\in\Lambda\hat{h}(\hat{y}_0(\pi)(t_1))+\int_t^{t_1}\zeta(s)^T\,\partial\hat{f}(s,\hat{y}_0(\pi)(s),\sigma_0\otimes\pi(s))\,ds\ \forall t\in T\Big\}$$

(b2) (*the maximum principles*).

(b2.1) (*the maximum principle in integral form*).

$$\int_{t_0}^{t_1}z(s)^Tf(s,y_0(s),(\sigma-\sigma_0)(s))\,ds+\int\omega(d\pi)\int_{t_0}^{t_1}\hat{z}(\pi)(s)^T\hat{f}(s,\hat{y}_0(\pi)(s),(\sigma-\sigma_0)(s),\pi(s))\,ds\geqq 0\quad\forall\sigma\in\mathcal{S}^{\#}$$

(b2.2) (*a "pointwise" maximum principle*). *If $R^\#(t)=R$ for all $t\in T$ then, for all $r\in R$,*

$$z(t)^T f(t, y_0(t), r)+\left[\int_W \hat{z}(\pi)(\cdot)^T\hat{f}(\cdot, \hat{y}_0(\pi)(\cdot), r, \pi(\cdot))\omega(d\pi)\right](t)$$

$$\geqq z(t)^T f(t, y_0(t), \sigma_0(t))$$

$$+\left[\int_W \hat{z}(\pi)(\cdot)^T\hat{f}(\cdot, \hat{y}_0(\pi)(\cdot), \sigma_0(\cdot), \pi(\cdot))\omega(d\pi)\right](t) \quad \mu\text{-a.e.}$$

(b2.3) (*a pointwise maximum principle*).

(i) *If the compact set on which the continuous function*

$$\pi\to\hat{h}(\hat{y}_0(\pi)(t_1)) : \mathscr{S}^\#_P\to\mathbb{R}$$

achieves its upper bound 0 *is at most denumerable then, for μ-almost all $t\in T$, $\sigma_0(t)$ has its support on the compact set on which the continuous function*

$$r\to z(t)^T f(t, y_0(t), r)+\int \hat{z}(\pi)(t)^T\hat{f}(t, \hat{y}_0(\pi)(t), r, \pi(t))\omega(d\pi)$$

achieves its minimum over $R^\#(t)$;

(ii) *if the function $\not{f}_2$ of assumption* (2.3.5) *is independent of r then, for μ-almost all $t\in T$, $\sigma_0(t)$ has its support on the compact set on which the continuous function*

$$r\to z(t)^T f(t, y_0(t), r)+\int \hat{z}(\pi)(t)^T\hat{\not{f}}_1(t, \hat{y}_0(\pi)(t), r)\omega(d\pi)$$

achieves its minimum over $R^\#(t)$.

(b3) (*the transversality condition*).

$$\hat{h}(\hat{y}_0(\pi^*)(t_1))=\max\{\hat{h}(\hat{y}_0(\pi)(t_1))\mid \pi\in\mathscr{S}^\#_P\}=0 \quad \textit{for } \omega\textit{-almost all } \pi^*.$$

THEOREM 2.5. *Let $\sigma_0\in\mathscr{S}^\#$, and let $\mathscr{S}'$ be an arbitrary denumerable subset of $\mathscr{S}^\#$ containing σ_0, with $\tilde{\mathscr{P}}$ defined accordingly. Assume that*

$$\hat{h}(\hat{y}(\sigma_0\otimes\pi)(t_1))\leqq 0 \quad \forall\pi\in\tilde{\mathscr{P}},$$

and let

$$y_0(t):=y(\sigma_0)(t), \qquad \hat{y}_0(\pi)(t):=\hat{y}(\sigma_0\otimes\pi)(t).$$

Let Assumption 2.3 *be satisfied and Λh and $\Lambda\hat{h}$ be derivate containers for h and $\hat{h}$. Then either*

(a) *there exists $\beta>0$ such that*

$$h(y_0(t_1))+\beta\bar{B}_m\subset\{h(y(\sigma)(t_1))\mid\sigma\in\operatorname{co}\mathscr{S}', \hat{h}(\hat{y}(\sigma\otimes\pi)(t_1))\leqq-\beta\ \forall\pi\in\tilde{\mathscr{P}}\}$$

or

(b) *there exist*

$$\ell^1\in\mathbb{R}^m, \quad \omega\in\operatorname{frm}^+(\tilde{\mathscr{P}}), \quad z\in AC(T,\mathbb{R}^n), \quad \hat{z}:\tilde{\mathscr{P}}\to AC(T,\mathbb{R}^{n+p})$$

such that

(b1)
$$0<|\ell^1|+\omega(\tilde{\mathcal{P}})\leqq 1$$

$$z(t)^T\in \ell^{1T}\Lambda h(y_0(t_1))+\int_t^{t_1} z(s)^T\,\partial f(s,y_0(s),\sigma_0(s))\,ds\quad \forall t\in T$$

$$\hat{z}(\pi)\in\overline{\mathrm{co}}\Big\{\zeta\in AC(T,\mathbb{R}^{n+p})\,|$$
$$\zeta(t)^T\in\Lambda\hat{h}(\hat{y}_0(\pi)(t_1))+\int_t^{t_1}\zeta(s)^T\,\partial\hat{f}(s,\hat{y}_0(\pi)(s),\sigma_0\otimes\pi(s))\,ds\ \forall t\in T\Big\}.$$

(b2) (*maximum principles*).

(b2.1) (*a weak maximum principle in integral form*).

$$\int_{t_0}^{t_1} z(s)^T f(s,y_0(s),(\sigma-\sigma_0)(s))\,ds$$
$$+\int\omega(d\pi)\int_{t_0}^{t_1}\hat{z}(\pi)(s)^T\hat{f}(s,\hat{y}_0(\pi)(s),(\sigma-\sigma_0)\otimes\pi(s))\,ds\geqq 0\quad\forall\sigma\in\mathcal{S}'.$$

(b2.2) (*a weak pointwise maximum principle*). *If $\mathcal{S}'$ is closed under concatenations with σ_0 on rational intervals, i.e., if, for every subinterval I of T with rational endpoints and every $\sigma\in\mathcal{S}'$, we have $\sigma'\in\mathcal{S}'$ whenever $\sigma'(t)=\sigma(t)$ for $t\in I$, $\sigma'(t)=\sigma_0(t)$, elsewhere, then, for all $\sigma\in\mathcal{S}'$,*

$$z(t)^T f(t,y_0(t),\sigma(t))+\left[\int_W \hat{z}(\pi)(\cdot)^T\hat{f}(\cdot,\hat{y}_0(\pi)(\cdot),\sigma\otimes\pi(\cdot))\omega(d\pi)\right](t)$$
$$\geqq z(t)^T f(t,y_0(t),\sigma_0(t))+\left[\int_W \hat{z}(\pi)(\cdot)^T\hat{f}(\cdot,\hat{y}_0(\pi)(\cdot),\sigma_0\otimes\pi(\cdot))\omega(d\pi)\right](t)$$
$$\mu\text{-a.e.}$$

(b3) (*the transversality condition*).
$$\hat{h}(\hat{y}_0(\pi^*)(t_1))=\max_{\pi\in\tilde{\mathcal{P}}}\hat{h}(\hat{y}_0(\pi)(t_1))=0\quad \textit{for } \omega\textit{-almost all } \pi^*.$$

Let
$$\mathcal{T}_i:=\Big\{\theta=(\theta^1,\cdots,\theta^i)\in\mathbb{R}^i\,|\,\theta^j\geqq 0,\ \sum_{j=1}^i\theta^j\leqq 1\Big\}.$$

If E and F are normed vector spaces, $A\subset E$ convex, $A^\circ\neq\emptyset$, $a\in A$, and $g:A\to F$, we define the (Fréchet) derivative $g'(a)$ (even if $a\in\partial A$) as the continuous linear operator on E to F such that
$$\lim|b-a|^{-1}[g(b)-g(a)-g'(a)(b-a)]=0\quad\text{as } b\to a,\quad b\in A\backslash\{a\}.$$

A basic tool in proving Theorems 2.4 and 2.5 will be a special case of the following general result that we expect to apply in other contexts as well. In proving Theorems 2.4 and 2.5 we will use Theorem 2.6 below with $\mathcal{U}=K$. However, in other applications that we have in mind, the set K may represent $\mathcal{S}^{\#}$ and the set $\mathcal{U}$ an "abundant" subset of $\mathcal{S}^{\#}$ such as the set of ordinary (unrelaxed) controls.

THEOREM 2.6. *Let K be a convex subset of a real vector space, $q_j\in K$ for $j=0,1,2,\cdots$, $\mathcal{Y}$ a normed vector space, C a convex subset of $\mathcal{Y}$ with a nonempty interior, $(\varphi,\Phi):K\to\mathbb{R}^m\times\mathcal{Y}$, $\Phi(q_0)\in C$, and e_j^i the jth column of the unit $i\times i$ matrix. Let $\mathcal{U}\subset K$ be such that, for each $i=1,2,\cdots$ and $\theta\in\mathcal{T}_i$, there exists a sequence $(u_k(\theta))$ in $\mathcal{U}$ such that, for each i,*

$$\lim_k(\varphi,\Phi)(u_k(\theta))=(\varphi,\Phi)\Big(q_0+\sum_{j=1}^i\theta^j(q_j-q_0)\Big)\quad\textit{uniformly on }\mathcal{T}_i$$

and

$$\theta \to (\varphi, \Phi)(u_k(\theta)) : \mathcal{T}_i \to \mathbb{R}^m \times \mathcal{Y} \quad \textit{is continuous for every } k.$$

Assume that there exist $\delta > 0$ *and functions* $(\varphi_i, \Phi_i) : K \to \mathbb{R}^m \times \mathcal{Y}$ *for* $i = 1, 2, \cdots$ *such that*

$$\lim_i (\varphi_i, \Phi_i) = (\varphi, \Phi) \quad \textit{uniformly}$$

and, for every $i = 1, 2, \cdots$, *the function*

$$\theta \to (g_i, G_i)(\theta) := (\varphi_i, \Phi_i)\left(q_0 + \sum_{j=1}^{i} \theta^j (q_j - q_0) \right) : \delta\mathcal{T}_i \to \mathbb{R}^m \times \mathcal{Y}$$

is continuously differentiable.

Then either

(a) *there exists* $\beta > 0$ *such that*

$$\varphi(q_0) + \beta \bar{B}_m \subset \{\varphi(u) \mid u \in \mathcal{U},\ \Phi(u) + \beta \bar{B}_{\mathcal{Y}} \subset C\}$$

or

(b) *there exist sequences* $(\ell_i) = (\ell_i^1, \ell_i^2)$ *in* $\mathbb{R}^m \times \mathcal{Y}^*$, (γ_i) *in* $(0, \infty)$, *and* (θ_i), *with* $|\ell_i| = |\ell_i^1| + |\ell_i^2| = 1$, $\lim_i \gamma_i = 0$, *and* $\theta_i \in \gamma_i \mathcal{T}_i$, *such that*

(b1) $\liminf_i \ell_i(g_i'(\theta_i), G_i'(\theta_i)) e_j^i \geqq 0$ *for all* $j = 1, 2, \cdots$,

and, for every weak cluster point* $\ell = (\ell^1, \ell^2)$ *of* (ℓ_i), *we have*

(b2) $0 < |\ell| = |\ell^1| + |\ell^2| \leqq 1$;

(b3) $\ell^2 \Phi(q_0) = \sup_{c \in C} \ell^2 c$.

3. Comments.

3.1. Maximum principles. Heuristically, one might "derive" the pointwise maximum principle 2.4(b2.3(i)) without the special assumption made there by interchanging the order of integration in the double integral in 2.4(b2.1) and then choosing various $\sigma \in \mathcal{S}^*$ to coincide with σ_0 except on small subintervals of T containing a given point t. This method—which works effectively for many other problems—fails in our case. The reason is that we have no way of showing, and no reason to expect, that the integrand of this double integral is $\omega \times \mu$-measurable. In fact, if we denote this integrand by $\chi(\pi, s)$, then we can think of the family $\{\chi(\pi, \cdot) \mid \pi \in \mathcal{P}\}$ as a stochastic process, and an analogous interchange of the order of integration in the context of stochastic processes appears to require rather strong assumptions (see, e.g., [3, Thm. 3A, p. 79]) that are not satisfied in our case. In our study of conflicting control problems with data that are C^1 with respect to the state variables [4, Chap. IX, § X.3.7] we were able to avoid this difficulty and to obtain improved results (both for relaxed and unrelaxed adverse controls) by proceeding essentially as follows. Let

$$\psi(\pi, t, r, r_P) := \hat{z}(\pi)(t)^T \hat{f}(t, \hat{y}(\sigma_0 \otimes \pi)(t), r, r_P),$$
$$\mathfrak{h}(\pi, t, r) = \max\{\psi(\pi, t, r, r_P) \mid r_P \in R_P^\#(t)\}.$$

We showed that relations 2.4(b3) and 2.5(b3) of Theorems 2.4 and 2.5, and specifically

$$\hat{h}(\hat{y}(\sigma_0 \otimes \pi^*)(t_1)) \geqq \hat{h}(\hat{y}(\sigma_0 \otimes \pi)(t_1)) \tag{3.1.1}$$

for ω-almost all π^* and all $\pi \in \mathcal{P}$, imply that, for μ-almost all t and $\sigma_0(t)$-almost all r, we have

$$\psi(\pi, t, r, \pi^*(t)) = \mathfrak{h}(\pi, t, r). \tag{3.1.2}$$

This last relation enabled us to replace the integrand of the double integral with $\mathfrak{h}(\pi, t, (\sigma - \sigma_0)(t))$ that is $\omega \times \mu$-measurable because $\pi \to \hat{y}(\sigma_0 \otimes \pi)(t_1)$ is continuous and $\pi \to \hat{z}(\pi)(\cdot): \mathcal{P} \to (C(T, \mathbb{R}^{n+p}), |\cdot|_{\sup})$ is ω-measurable so that $(\pi, s) \to \hat{z}(\pi)(s)$ is $\omega \times \mu$-measurable.

The reason that the above method is effective in the C^1 case is because $\hat{z}(\pi)$ is then uniquely defined by the differential equation to which the differential inclusions of 2.4(b1) and 2.5(b1) reduce themselves. Relation (3.1.1) then implies that

$$\int_{t_0}^{t_1} \psi(\pi, s, \sigma_0(s), \pi'(s) - \pi^*(s))\, ds,$$

which is the directional derivative of $\pi \to \hat{h}(\hat{y}(\sigma_0 \otimes \pi)(t_1))$ at π^* in the direction of $\pi' - \pi^*$, must be nonpositive for all $\pi' \in \mathcal{P}$, and this yields relation (3.1.2). Unfortunately, in our present context, the function $\hat{z}(\pi)$ that "corresponds" to relation (3.1.1) and the function $\hat{z}(\pi)$ that appears in 2.4(b2), respectively, 2.5(b2) may be different solutions of the inclusion in (b1).

The "pointwise" maximum principles 2.4(b2.2) and 2.4(b2.3(i)) appear to be both theoretically and computationally much weaker than the corresponding C^1 results involving the function $\mathfrak{h}$. It remains to be seen whether, and to what extent, these results can be strengthened. The maximum principles of Theorem 2.5 are much weaker than the corresponding relations 2.4(b2.1) and 2.4(b2.2) of Theorem 2.4 and, for this reason, Theorem 2.5 must be viewed as a preliminary result marking only the first step in the study of nonsmooth problems with hyperrelaxed adverse controls. On the other hand, Theorem 2.5 "comes cheaply" because it is derived along with Theorem 2.4 without causing any complications or requiring any additions.

3.2. Assumption (2.3.5). For all $\pi \in \mathcal{P}$, let

$$Z(\pi) = \Big\{ \zeta \in AC(T, \mathbb{R}^{n+p}) \,|\, \zeta(t)^T \in \Lambda \hat{h}(\hat{y}_0(\pi)(t_1)) + \int_t^{t_1} \zeta(s)^T\, \partial \hat{f}(s, \hat{y}_0(\pi)(s), \sigma_0 \otimes \pi(s))\, ds \ \forall t \in T \Big\}.$$

Assumption (2.3.5) is explicitly used only in Lemma 4.4 below to prove that the set-valued function $\pi \to Z(\pi)$ has a closed graph. The latter property is later used to prove the relation $\hat{z}(\pi) \in \overline{\text{co}}\, Z(\pi)$ ω-almost everywhere of statements 2.4(b1) and 2.5(b1). In view of the transversality relations 2.4(b3) and 2.5(b3), this last relation involves only the adverse controls π in the set $\mathcal{H}$ of those that maximize $\hat{h}(\hat{y}(\sigma_0 \otimes \pi)(t_1))$. For a given σ_0, the determination of $\mathcal{H}$ represents an unrestricted optimal control problem (at least in the case of relaxed adverse controls), a problem that is reasonably well understood and often computationally manageable. On the other hand, without Assumption (2.3.5), we can only assert that $\hat{z}(\pi) \in \overline{\text{co}}\, \mathfrak{Z}(\pi)$ ω-almost everywhere, where $\mathfrak{Z}(\cdot)$ is the set-valued mapping whose graph is the closure of graph $(Z(\cdot))$. The use of $\mathfrak{Z}(\cdot)$ would generally require the determination of lim-sup $Z(\pi)$ as π approaches $\mathcal{H}$ (a problem for which we know no practicable methods of attack) except for special problems, such as evasion problems [4, § IX.3], where $\hat{f}$ has a special structure.

4. Proofs. We denote by $\mathcal{Y}$ a normed vector space. If A and B are subsets of a real vector space, we write

$$A \ominus B := \{z \in A \,|\, z + B \subset A\}.$$

LEMMA 4.1. *If $A \subset \mathcal{Y}$ and $W \subset \mathbb{R}^m \times \mathcal{Y}$ are convex, $A^\circ \neq \varnothing$, $z \in \mathbb{R}^m$, and $W \cap \{z\} \times A = \varnothing$ then there exists $\ell = (\ell^1, \ell^2) \in \mathbb{R}^m \times \mathcal{Y}^*$ such that*

$$|\ell| = 1, \qquad \ell w \geqq \ell^1 z + \ell^2 a \quad \forall w \in W, \quad a \in A.$$

Proof. Let $V:=\{v\,|\,(z,v)\in W\}$. Then $V\cap A=\emptyset$ and there exists $\ell'^2\in\mathcal{Y}^*$ such that $\ell'^2\neq 0$ and

$$\ell'^2 v \geqq \ell'^2 a \quad \forall v\in V, \quad a\in A.$$

Thus

$$\{(\ell'^2 w^2, w^1)\,|\,(w^1,w^2)\in W\} \quad \text{and} \quad \{(\ell'^2 a-\alpha, z)\,|\,\alpha>0,\ a\in A\}$$

are disjoint and convex subsets of $\mathbb{R}\times\mathbb{R}^m$ and there exists $(\ell^0,\ell^1)\neq 0$ in $\mathbb{R}\times\mathbb{R}^m$ such that

$$\ell^0\ell'^2 w^2+\ell^1 w^1 \geqq \ell^0\ell'^2 a-\ell^0\alpha+\ell^1 z \quad \forall\alpha>0, \quad (w_1,w_2)\in W, \quad a\in A.$$

Set $\ell^2=\ell^0\ell'^2$. If $\ell^1=0$ then $\ell^0\neq 0$, hence $\ell^2\neq 0$. Thus $\ell:=(\ell^1,\ell^2)\neq 0$ and

$$\ell w\geqq \ell^1 z+\ell^2 a \quad \forall w\in W, \quad a\in A.$$

We may assume that $|\ell|=1$, otherwise replacing ℓ with $\ell/|\ell|$. □

LEMMA 4.2. *Let W be a nonempty convex subset of $\mathbb{R}^m\times\mathcal{Y}$, $r>\gamma>0$, C_0 a convex subset of $\mathcal{Y}$ containing $y+rB_{\mathcal{Y}}$, $0\in W$, $0\in\bar{C}_0$, and $C_\gamma:=C_0\ominus\gamma\bar{B}_{\mathcal{Y}}$. Then either*

(a) $\gamma\bar{B}_m\subset\{w^1\,|\,(w^1,w^2)\in W,\ w^2\in C_\gamma\}$

or

(b) *there exists $\ell=(\ell^1,\ell^2)\in\mathbb{R}^m\times\mathcal{Y}^*$ such that $|\ell|=|\ell^1|+|\ell^2|=1$ and*

$$\ell w\geqq-\gamma(1+|y|/r) \quad \forall w\in W, \qquad \ell^2 c_\gamma\leqq\gamma \quad \forall c_\gamma\in C_\gamma.$$

Proof. Since $r>\gamma$, the set C_γ° is nonempty. If (a) does not hold then there exists a point $z\in\gamma\bar{B}_m$ such that $W\cap\{z\}\times C_\gamma=\emptyset$. Then, by Lemma 4.1, there exists $\ell\in\mathbb{R}^m\times\mathcal{Y}^*$ such that $|\ell|=1$ and

$$\ell^1 z+\ell^2 c_\gamma\leqq\ell w \quad \forall c_\gamma\in C_\gamma, \qquad w\in W.$$

Since $0\in W$, this implies that

$$\ell^2 c_\gamma\leqq-\ell^1 z\leqq|\ell|\,|z|\leqq\gamma \quad \forall c_\gamma\in C_\gamma.$$

Since $(\gamma/r)y\in C_\gamma$, we have

$$\ell w\geqq\ell^1 z+\frac{\gamma}{r}\ell^2 y\geqq-|\ell^1|\gamma-\frac{\gamma}{r}|\ell^2|\,|y|\geqq-\gamma(1+|y|/r) \quad \forall w\in W. \qquad \square$$

4.3. Proof of Theorem 2.6. Let

$$A_i(\theta,\alpha):=\{g_i'(\theta)\omega\,|\,\omega\in\mathcal{T}_i,\ G_i'(\theta)\omega+\alpha\bar{B}_{\mathcal{Y}}\subset C-\Phi(q_0)\}$$

$$\gamma_i':=\sup\{\alpha\in[0,\delta]\,|\,\alpha\bar{B}_m\subset A_i(\theta,\alpha)\ \forall\theta\in\alpha\mathcal{T}_i\}.$$

Then either $\gamma=\frac{1}{2}\limsup_i\gamma_i'>0$, and then

(i) there exist $\gamma\in(0,\delta]$ and $\mathcal{I}_1\subset(1,2,\cdots)$ such that

$$\gamma\bar{B}_m\subset A_i(\theta,\gamma) \quad \forall i\in\mathcal{I}_1, \qquad \theta\in\gamma\mathcal{T}_i$$

or $\lim_i\gamma_i'=0$, and then

(ii) there exist i_0, $\gamma_i=\gamma_i'+1/i>0$, and $\theta_i\in\gamma_i\mathcal{T}_i$ such that $\lim_i\gamma_i=0$ and $\gamma_i\bar{B}_m$ is not contained in $A_i(\theta_i,\gamma_i)$ for $i\geqq i_0$.

First assume that (i) holds, and let $(u_k(\theta))$ be as defined in the statement of the theorem. Let $I, N \in \{1, 2, \cdots\}$ be sufficiently large so that

$$\sup_{\theta \in \gamma \mathcal{T}_I} |g_I(\theta) - \varphi(u_N(\theta))| \leqq \gamma^2/32, \tag{4.3.1}$$

$$\sup_{\theta \in \gamma \mathcal{T}_I} |G_I(\theta) - \Phi(u_N(\theta))| \leqq \gamma^2/128, \tag{4.3.2}$$

$$|G_I(0) - \Phi(q_0)| \leqq \gamma^2/128, \tag{4.3.3}$$

and let

$$G_I^{\#}(\theta) := G_I(\theta) + [\Phi(q_0) - G_I(0)].$$

Since $\Phi(q_0) \in C$, we have $G_I^{\#}(0) \in C$. Furthermore, by (4.3.2) and (4.3.3),

$$\sup_{\theta \in \gamma \mathcal{T}_I} |G_I^{\#}(\theta) - \Phi(u_N(\theta))| \leqq \gamma^2/64. \tag{4.3.4}$$

Let $\beta = \gamma^2/64$. Since $G_I^{\#\prime}(\theta) = G_I'(\theta)$ for all $\theta \in \delta \mathcal{T}_I$, it follows from (4.3.1), (4.3.4), and [7, Thm. 3.4, p. 848], that

$$\begin{aligned}\varphi(q_0) + \beta \bar{B}_m &\subset \{\varphi(u_N(\theta)) \mid \theta \in \gamma \mathcal{T}_I,\ \Phi(u_N(\theta)) + \beta \bar{B}_{\mathcal{Y}} \subset C\} \\ &\subset \{\varphi(u) \mid u \in \mathcal{U},\ \Phi(u) + \beta \bar{B}_{\mathcal{Y}} \subset C\},\end{aligned}$$

showing that alternative (a) is valid.

Next assume that (ii) holds. Since $C^\circ \neq \emptyset$, there exist $y \in C$ and $r > 0$ such that $y + r\bar{B}_{\mathcal{Y}} \subset C$, and we may assume that $r > \gamma_i$ for all sufficiently large i, say $i \geqq i_1$. By Lemma 4.2, with

$$W := \{(g_i'(\theta_i), G_i'(\theta_i))\omega \mid \omega \in \mathcal{T}_i\}, \qquad C_0 := C - \Phi(q_0),$$

for each $i \geqq i_1$ there exists $\ell_i = (\ell_i^1, \ell_i^2) \in \mathbb{R}^m \times \mathcal{Y}^*$ such that $|\ell_i| = 1$,

$$\ell_i^1 g_i'(\theta_i)\omega + \ell_i^2 G_i'(\theta_i)\omega \geqq -\gamma_i(1 + |y|/r) \quad \forall \omega \in \mathcal{T}_i, \tag{4.3.5}$$

$$\ell_i^2 c_i \leqq \gamma_i \quad \forall c_i \in C - \Phi(q_0) \ominus \gamma_i \bar{B}_{\mathcal{Y}}. \tag{4.3.6}$$

Relation (b1) now follows directly from (4.3.5).

Let $\ell = (\ell^1, \ell^2)$ be the weak* limit of some subsequence $(\ell_i)_{i \in \mathcal{I}}$ of (ℓ_i), with $\mathcal{I} \subset (i_1, i_1 + 1, \cdots)$, and let $c \in C^\circ$. Then, for all sufficiently large $i \in \mathcal{I}$,

$$c - \Phi(q_0) \in C - \Phi(q_0) \ominus \gamma_i \bar{B}_{\mathcal{Y}}$$

and by (4.3.6) $\ell^2(c - \Phi(q_0)) \leqq 0$. Since ℓ^2 is continuous, this implies that

$$\ell^2(c - \Phi(q_0)) \leqq 0 \quad \forall c \in C,$$

thus proving (b3). Thus it remains to prove that $0 < |\ell| \leqq 1$. The inequality $|\ell| \leqq 1$ follows from $|\ell_i| = 1$ for all i. The inequality $|\ell| > 0$ follows directly from (4.3.6) and the assumption that $C^\circ \neq \emptyset$. Indeed, there exist y_0 and $r_0 > 0$ such that

$$y_0 + r_0 \bar{B}_{\mathcal{Y}} \subset C - \Phi(q_0) \ominus \gamma_i \bar{B}_{\mathcal{Y}}$$

for all large $i \in \mathcal{I}$. It follows by (4.3.6) that

$$\ell_i^2 y_0 + r_0 \ell_i^2 z \leqq \gamma_i \quad \forall z \in \bar{B}_{\mathcal{Y}};$$

hence

$$\ell_i^2 y_0 \leqq \gamma_i - r_0 |\ell_i^2| = \gamma_i - r_0(1 - |\ell_i^1|)$$

and therefore

$$\ell^2 y_0 \leqq -r_0(1-|\ell^1|),$$

thus proving that $\ell = (\ell^1, \ell^2) \neq 0$. □

We will henceforth use the notation of § 2 (and, in particular, the symbol $\mathcal{P}$ to denote either $\mathcal{S}^\#_P$ or $\tilde{\mathcal{P}}$) and the known result (proven in the arguments of [4, Thm. VI.1.1]) that the functions

$$\sigma \to y(\sigma) : \mathcal{S}^\# \to C(T, \mathbb{R}^n) \quad \text{and} \quad \hat{\sigma} \to \hat{y}(\hat{\sigma}) : \hat{\mathcal{S}}^\# \to C(T, \mathbb{R}^{n+p})$$

are continuous. We also observe that the μ-integrable function ψ of assumption (2.3.3) may be assumed to be a constant. This can be accomplished by replacing the independent variable t in the differential equation

$$\hat{y}(t) = \int_{t_0}^{t} \hat{f}(s, \hat{y}(s), \sigma \otimes \pi(s))\, ds \quad \forall t \in T$$

by $\tau := \int_{t_0}^{t} [1+\psi(s)]\, ds$ that effectively divides $\hat{f}$ by $1+\psi(t)$ and does not affect the statement of Theorems 2.4 and 2.5. Furthermore, since Assumption 2.3 guarantees that all the solutions of the above equations are uniformly bounded, we denote by c the common bound and Lipschitz constant of $f(t, \cdot, r)$, $\hat{f}(t, \cdot, r, r_P)$, $\hat{f}_1(t, \cdot, r)$, $\hat{f}_2(t, \cdot, r, r_P)$, h, $\hat{h}$, $y(\sigma)$, and $\hat{y}(\sigma \otimes \pi)$.

LEMMA 4.4. *Let $c^* := c \exp(c(t_1 - t_0))$ and X be the compact subset of $C(T, \mathbb{R}^{n+p})$ whose elements have c^* as a common bound and a common Lipschitz constant. For each $\pi \in \mathcal{P}$, let $Z(\pi)$ be the collection of all $\zeta \in AC(T, \mathbb{R}^{n+p})$ satisfying*

$$\zeta(t)^T \in \Lambda \hat{h}(\hat{y}_0(\pi)(t_1)) + \int_t^{t_1} \zeta(s)^T\, \partial \hat{f}(s, \hat{y}_0(\pi)(s), \sigma_0 \otimes \pi(s))\, ds \quad \forall t \in T.$$

Then $Z(\pi) \subset X$ for all $\pi \in \mathcal{P}$ and the function

$$\pi \to Z(\pi) : \mathcal{P} \to 2^{AC(T, \mathbb{R}^{n+p})}$$

has a closed graph.

Proof. Let $\lim_k \pi_k = \pi^*$ in $\mathcal{P}$. Since $\partial \hat{f}$ and $\Lambda \hat{h}$ are bounded by c for all $\pi \in \mathcal{P}$ and $\partial \hat{f}$ has values that are compact and convex, the sets $Z(\pi)$ are compact and contained in X. If, for each $k = 1, 2, \cdots$, we choose $\zeta_k \in Z(\pi_k)$ and set

$$\hat{y}_k := \hat{y}(\sigma_0 \otimes \pi_k) = \hat{y}_0(\pi_k), \qquad \hat{\sigma}_k := \sigma_0 \otimes \pi_k$$

then, by Assumption (2.3.5),

$$\partial \hat{f}(t, \hat{y}_k(t), \hat{\sigma}_k(t)) = \partial \hat{f}_1(t, \hat{y}_k(t), \sigma_0(t)) + \mathcal{D}_2 \hat{f}_2(t, \hat{y}_k(t), \hat{\sigma}_k(t))$$

and there exist μ-measurable selections M_k of

$$t \to \partial \hat{f}_1(t, \hat{y}_k(t), \sigma_0(t))$$

such that, for all $t \in T$,

$$\zeta_k(t)^T = \zeta_k(t_1)^T + \int_t^{t_1} \zeta_k(s)^T [M_k(s) + \mathcal{D}_2 \hat{f}_2(s, \hat{y}_k(s), \hat{\sigma}_k(s))]\, ds. \tag{4.4.1}$$

Now assume, by way of contradiction, that there exist $\mathcal{K} \subset (1, 2, \cdots)$ and $\varepsilon_0 > 0$ such that

$$\operatorname{dist}(\zeta_k, Z(\pi^*)) := \inf\{|\zeta_k - x|_{\sup} \mid x \in Z(\pi^*)\} > \varepsilon_0 \quad \forall k \in \mathcal{K}. \tag{4.4.2}$$

We may choose a subsequence $\mathscr{K}_1$ of $\mathscr{K}$ such that the functions ζ_k converge uniformly to some ζ^* as k ranges over $\mathscr{K}_1$. Since $\lim_k \pi_k = \pi^*$ in $\mathscr{P}$, hence $\lim_k \sigma_0 \otimes \pi_k = \sigma_0 \otimes \pi^*$ in $\hat{\mathscr{S}}^*$, we have

$$\lim_k \hat{y}(\sigma_0 \otimes \pi_k) = \hat{y}(\sigma_0 \otimes \pi^*) \text{ in } C(T, \mathbb{R}^{n+p}).$$

Let $\bar{B}_{n+p,n+p}$ denote $\bar{B}_{\mathscr{Y}}$ for $\mathscr{Y} = \mathscr{L}(\mathbb{R}^{n+p}, \mathbb{R}^{n+p})$, $c_1 = 2c(t_1 - t_0)$,

$$c' = (3cc^* + 1)[1 + c_1 \exp(c_1)], \qquad 0 < \eta < \varepsilon_0/c', \quad \hat{y}^* = \hat{y}(\sigma_0 \otimes \pi^*)$$

and let $\mathscr{K}_2 \subset \mathscr{K}_1$ be such that, for all $k \in \mathscr{K}_2$, $r \in R$, $r_P \in R_P$, $t \in T$,

$$\begin{gathered} |\mathscr{D}_2 \hat{f}_2(t, \hat{y}_k(t), r, r_P) - \mathscr{D}_2 \hat{f}_2(t, \hat{y}^*(t), r, r_P)| \leqq \eta \\ \partial \hat{f}_1(t, \hat{y}_k(t), \sigma_0(t)) \subset \partial \hat{f}_1(t, \hat{y}^*(t), \sigma_0(t)) + \eta \bar{B}_{n+p,n+p} \\ |\zeta_k - \zeta^*|_{\sup} \leqq \eta, \qquad \zeta_k(t_1) \in \Lambda \hat{h}(\hat{y}^*(t_1)) + \eta \bar{B}_p. \end{gathered} \tag{4.4.3}$$

Then, for $k \in \mathscr{K}_2$,

$$\begin{aligned} \lim_{j \in \mathscr{K}_2} \int_t^{t_1} & \zeta_j(s)^T \mathscr{D}_2 \hat{f}_2(s, \hat{y}_j(s), \hat{\sigma}_j(s))\, ds \\ &= \int_t^{t_1} \zeta^*(s)^T \mathscr{D}_2 \hat{f}_2(s, \hat{y}^*(s), \sigma_0 \otimes \pi^*(s))\, ds \\ &\in \int_t^{t_1} \zeta_k(s)^T \mathscr{D}_2 \hat{f}_2(s, \hat{y}^*(s), \sigma_0 \otimes \pi^*(s))\, ds + c\eta \bar{B}_p \end{aligned}$$

and therefore

$$\begin{aligned} \int_t^{t_1} & \zeta_k(s)^T \mathscr{D}_2 \hat{f}_2(s, \hat{y}_k(s), \sigma_0 \otimes \pi_k(s))\, ds \\ &\in \int_t^{t_1} \zeta_k(s)^T \mathscr{D}_2 \hat{f}_2(s, \hat{y}^*(s), \sigma_0 \otimes \pi^*(s))\, ds + 2c\eta \bar{B}_p \end{aligned} \tag{4.4.4}$$

for all sufficiently large $k \in \mathscr{K}_2$, say for $k \in \mathscr{K}_3$.

Let $M_k^*(t)$ be the closest point in the compact convex set $\partial \hat{f}_1(t, \hat{y}^*(t), \sigma_0(t))$ to $M_k(t)$ and $\mathfrak{z}_k$ a closest point in $\Lambda \hat{h}(\hat{y}^*(t_1))$ to $\zeta_k(t_1)$. The point $M_k^*(t)$ is unique because we have chosen a strictly convex norm for $\mathscr{L}(\mathbb{R}^{n+p}, \mathbb{R}^{n+p})$. By Lusin's theorem, the μ-measurable functions $t \to M_k(t)$ and $t \to \partial \hat{f}_1(t, \hat{y}^*(t), \sigma_0(t))$ have continuous restrictions to sets of nearly full μ-measure, and $M_k^*(\cdot)$ is continuous when restricted to these sets. Thus $M_k^*(\cdot)$ is μ-measurable and, by (4.4.1), (4.4.3), and (4.4.4), for all $k \in \mathscr{K}_3$ and $t \in T$, we have

$$\zeta_k(t)^T = \mathfrak{z}_k + \int_t^{t_1} \zeta_k(s)^T [M_k^*(s) + \mathscr{D}_2 \hat{f}_2(s, \hat{y}^*(s), \sigma_0 \otimes \pi^*(s))]\, ds + e_k(t)^T,$$

where $|e_k(t)| \leqq (3cc^* + 1)\eta$. If we denote by $\zeta_k^*(\cdot)$ the solution ζ of

$$\zeta(t)^T = \mathfrak{z}_k + \int_t^{t_1} \zeta(s)^T [M_k^*(s) + \mathscr{D}_2 \hat{f}_2(s, \hat{y}^*(s), \sigma_0 \otimes \pi^*(s))]\, ds,$$

then we observe that $\zeta_k^* \in Z(\pi^*)$ for all $k \in \mathscr{K}_3$ and that $\Delta_k := \zeta_k - \zeta_k^*$ satisfies

$$\Delta_k(t) = \int_t^{t_1} \Delta_k(s)^T [M_k^*(s) + \mathscr{D}_2 \hat{f}_2(s, \hat{y}^*(s), \sigma_0 \otimes \pi^*(s))]\, ds + e_k(t).$$

Thus, by Gronwall's inequality,

$$|\Delta_k(t)| \leqq (3cc^*+1)[1+2c(t_1-t_0)\exp(2c(t_1-t_0))]\eta = c'\eta$$

and therefore

$$\zeta_k \in Z(\pi^*) + c'\eta \bar{B}_{C(T,\mathbb{R}^{n+p})} \subset Z(\pi^*) + \varepsilon_0 \bar{B}_{C(T,\mathbb{R}^{n+p})} \quad \forall k \in \mathscr{K}_3.$$

Thus $\operatorname{dist}(\zeta_k, Z(\pi^*)) \leqq \varepsilon_0$, which contradicts (4.4.2).

We conclude that the function $\pi \to Z(\pi) : \mathscr{P} \to 2^{AC(T,\mathbb{R}^{n+p})}$ is upper semicontinuous and has compact values. Thus, by Berge's theorem, it has a closed graph. □

The following lemma is partly patterned on Lemma 3.2 of [5, p. 30].

LEMMA 4.5. *Let* $\mathscr{X} = C(T, \mathbb{R}^{n+p})$, $\omega_i \in \operatorname{frm}^+(\mathscr{P})$, $\omega_i(\mathscr{P}) \leqq 1$, $\lim_i \omega_i = \omega$ *weakly*, X *and* $Z(\pi)$ *be as defined in Lemma* 4.4, $\hat{z}_i : \mathscr{P} \to X$ ω_i*-measurable for all* $i = 1, 2, \cdots$, *and assume that for every* $\varepsilon > 0$ *there exists* $i_0(\varepsilon)$ *such that*

$$\hat{z}_i(\pi) \in \Gamma_\varepsilon(\pi) := (Z(\pi) + \varepsilon \bar{B}_{\mathscr{X}}) \cap X \quad \forall i \geqq i_0(\varepsilon), \qquad \pi \in \mathscr{P}.$$

Then there exist an ω*-measurable* $\hat{z} : \mathscr{P} \to C(T, \mathbb{R}^{n+p})$ *and* $\mathscr{I} \subset (1, 2, \cdots)$ *such that*

$$\hat{z}(\pi) \in \overline{\operatorname{co}}\, Z(\pi) \quad \forall \pi \in \mathscr{P}$$

and

$$\lim_{i \in \mathscr{I}} \int \omega_i(d\pi) \int_{t_0}^{t_1} \hat{z}_i(\pi)(s)^T \chi(s, \pi)\, ds = \int \omega(d\pi) \int_{t_0}^{t_1} \hat{z}(\pi)(s)^T \chi(s, \pi)\, ds$$

for all bounded $\chi : T \times \mathscr{P} \to \mathbb{R}^{n+p}$ *such that* $s \to \chi(s, \pi)$ *is* μ*-measurable for each* $\pi \in \mathscr{P}$ *and the function*

$$\pi \to \int_{t_0}^{t_1} x(s)^T \chi(s, \pi)\, ds : \mathscr{P} \to \mathbb{R}$$

is continuous for each $x \in X$.

Proof. For $i = 1, 2, \cdots$, let

$$\mathscr{L}_i(\varphi) = \int \varphi(\pi, \hat{z}_i(\pi))\omega_i(d\pi) \quad \forall \varphi \in C(\mathscr{P} \times X).$$

Then there exists $\nu_i \in \operatorname{frm}^+(\mathscr{P} \times X)$ such that

$$\text{(4.5.1)} \qquad \mathscr{L}_i(\varphi) = \int \varphi(\pi, x)\nu_i(d(\pi, x)), \qquad |\mathscr{L}_i| = \nu_i(\mathscr{P} \times X) = \omega_i(\mathscr{P}) \leqq 1$$

for all $\varphi \in C(\mathscr{P} \times X)$, and there exist $\mathscr{I} \subset (1, 2, \cdots)$ and $\nu \in \operatorname{frm}^+(\mathscr{P} \times X)$ such that

$$\lim_{i \in \mathscr{I}} \nu_i = \nu \quad \text{weakly.}$$

Let n_ω denote the (conventional) norm of $L^1(\omega, C(\mathscr{P}))$. Then, for all $\varphi \in C(\mathscr{P} \times X) = C(\mathscr{P}, C(X))$, we have

$$\left| \int \varphi(\pi, x)\nu(d(\pi, x)) \right| \leqq \int |\varphi(\pi, \cdot)|_{\sup} \nu(d(\pi, x))$$

$$= \lim_{i \in \mathscr{I}} \int |\varphi(\pi, \cdot)|_{\sup} \nu_i(d(\pi, x)) = \lim_{i \in \mathscr{I}} \int |\varphi(\pi, \cdot)|_{\sup} \omega_i(d\pi)$$

$$= \int |\varphi(\pi, \cdot)|_{\sup} \omega(d\pi) = n_\omega(\varphi).$$

Thus ν is a continuous linear functional on the normed vector space $(C(\mathcal{P}, C(X)), n_\omega)$ and, by the Hahn-Banach theorem, can be extended to $(L^1(\omega, C(X)), n_\omega)$. It follows, by a variant of the Dunford-Pettis theorem (the proof being the same as in [4, Thm. IV.1.8]), that there exists an ω-measurable $\lambda : \mathcal{P} \to \operatorname{frm}(X)$ such that $\operatorname{ess\,sup}_{t\in T} |\lambda(t)|(X) < \infty$ and

$$(4.5.2) \int \varphi(\pi, x)\nu(d(\pi, x)) = \int \omega(d\pi) \int \varphi(\pi, x)\lambda(\pi)(dx) \quad \forall \varphi \in L^1(\omega, C(X)).$$

We verify that $\lambda(\pi)$ is a nonnegative measure ω-almost everywhere.

Let $\varepsilon > 0$. By Lemma 4.4, Graph (Z) is closed and therefore Graph (Γ_ε) is closed. Since $\hat{z}_i(\pi) \in \Gamma_\varepsilon(\pi)$ for all $\pi \in \mathcal{P}$ and all sufficiently large i, relation (4.5.1) implies that ν_i has its support on Graph (Γ_ε) for sufficiently large i and therefore ν has its support on Graph (Γ_ε). It follows thus from relation (4.5.2) that $\lambda(\pi)$ is supported ω-almost everywhere on $\Gamma_\varepsilon(\pi)$. Therefore, by (4.5.1) and (4.5.2), for every $u \in C(\mathcal{P})$,

$$\begin{aligned}\lim_{i\in\mathcal{I}} \mathcal{L}_i(u) = \int u(\pi)\omega(d\pi) &= \int u(\pi)\nu(d(\pi, x)) \\ &= \int \omega(d\pi) \int u(\pi)\lambda(\pi)(dx) \\ &= \int u(\pi)\lambda(\pi)(\Gamma_\varepsilon(\pi))\omega(d\pi);\end{aligned}$$

hence $\lambda(\pi)(\Gamma_\varepsilon(\pi)) = 1$ ω-almost everywhere.

Now set

$$\hat{z}(\pi) := \int x\lambda(\pi)(dx) \quad \forall \pi \in \mathcal{P}$$

and let $\chi : T \times \mathcal{P} \to \mathbb{R}^{n+p}$ be as described in the statement of the lemma. Then $\pi \to \hat{z}(\pi)$ is ω-measurable and

$$\hat{z}(\pi) \in \overline{\text{co}}\, \Gamma_\varepsilon(\pi) := \overline{\text{co}}\, [(Z(\pi) + \varepsilon \bar{B}_{\mathcal{X}}) \cap X] \quad \forall \varepsilon > 0;$$

hence $\hat{z}(\pi) \in \overline{\text{co}}\, Z(\pi)$. Furthermore, the function

$$(\pi, x) \to \varphi(\pi, x) = \int_{t_0}^{t_1} x(s)^T \chi(s, \pi)\, ds$$

being continuous in π for each x and continuous in x uniformly for all π, belongs to $C(\mathcal{P} \times X)$. Thus, by (4.5.1) and (4.5.2),

$$\begin{aligned}\mathcal{H}(\chi) := \lim_{i\in\mathcal{I}} \int \omega_i(d\pi) \int_{t_0}^{t_1} \hat{z}_i(\pi)(s)^T \chi(s, \pi)\, ds \\ = \int \nu(d(\pi, x)) \int_{t_0}^{t_1} x(s)^T \chi(s, \pi)\, ds \\ = \int \omega(d\pi) \int \left[\int_{t_0}^{t_1} x(s)^T \chi(s, \pi)\, ds\right] \lambda(\pi)(dx).\end{aligned}$$

We observe that, for every $\pi \in \mathcal{P}$, the function $s \to x(s)^T \chi(s, \pi)$ is μ-integrable for all $x \in X$ and the function $x \to x(s)^T \chi(s, \pi)$ is continuous on X uniformly for all s and

π. Thus the function $(s, x)\to x(s)^T\chi(s, \pi)$ is $\mu\times\lambda(\pi)$-integrable and, by Fubini's theorem,

$$\int\left[\int_{t_0}^{t_1} x(s)^T\chi(s, \pi)\, ds\right]\lambda(\pi)(dx)=\int_{t_0}^{t_1}\left[\int x(s)^T\lambda(\pi)(dx)\right]\chi(s, \pi)\, ds$$

$$=\int_{t_0}^{t_1}\left[\int x\lambda(\pi)(dx)\right](s)^T\chi(s, \pi)\, ds$$

$$=\int_{t_0}^{t_1}\hat{z}(\pi)(s)^T\chi(s, \pi)\, ds.$$

Thus $\mathcal{H}(\chi)=\int\omega(d\pi)\int_{t_0}^{t_1}\hat{z}(\pi)(s)^T\chi(s, \pi)\, ds.$ □

LEMMA 4.6. *Let ω and $\hat{z}$ be as in Lemma* 4.5 *and*

$$\hat{\mathcal{A}}_j(\pi, t):=\hat{z}(\pi)(t)^T\hat{f}(t, \hat{y}_0(\pi)(t), (\sigma_j-\sigma_0)\otimes\pi(t)).$$

Then

$$\int\omega(d\pi)\int_{t_0}^{t_1}\hat{\mathcal{A}}_j(\pi, s)\, ds=\int_{t_0}^{t_1}\left[\int_w\hat{\mathcal{A}}_j(\pi, \cdot)\omega(d\pi)\right](s)\, ds\quad\forall j=1, 2, \cdots.$$

Proof. We first prove that the bounded function

$$\pi\to\hat{\mathcal{A}}_j(\pi, \cdot):\mathcal{P}\to(L^2(\mu, \mathbb{R}), |\cdot|_w)$$

is ω-measurable and therefore also ω-integrable. Since

$$\pi\to\hat{z}(\pi):\mathcal{P}\to(C(T, \mathbb{R}^{n+p}), |\cdot|_{\sup})$$

is ω-measurable, for every $\varepsilon>0$ there exists a closed set $F_\varepsilon\subset\mathcal{P}$ such that $\omega(\mathcal{P}\backslash F_\varepsilon)<\varepsilon$ and $\hat{z}|_{F_\varepsilon}$ is continuous, so that

$$\lim_i\hat{z}(\pi_i)(t)=\hat{z}(\pi)(t)\quad\text{uniformly for } t\in T$$

if $\pi, \pi_i\in F_\varepsilon$ and $\lim_i\pi_i=\pi$. The bounded function

$$\pi\to\hat{f}(\cdot, \hat{y}_0(\pi)(\cdot), (\sigma_j-\sigma_0)\otimes\pi(\cdot)):\mathcal{P}\to(L^2(\mu, \mathbb{R}), |\cdot|_w)$$

is continuous because

$$\pi\to\hat{y}_0(\pi):\mathcal{P}\to(C(T, \mathbb{R}^{n+p}), |\cdot|_{\sup})$$

is continuous and therefore

$$\pi\to\int_{t_0}^{t_1} g(s)^T\hat{f}(s, \hat{y}_0(\pi)(s), (\sigma_j-\sigma_0)\otimes\pi(s))\, ds$$

is continuous for every $g\in L^2(\mu, \mathbb{R}^{n+p})$. Therefore, if $\lim_i\pi_i=\pi$ in F_ε then

$$\lim_i\int_{t_0}^{t_1} g(s)\mathcal{A}_j(\pi_i, s)\, ds=\int_{t_0}^{t_1} g(s)\mathcal{A}_j(\pi, s)\, ds\quad\forall g\in L^2(\mu, \mathbb{R}).$$

Thus $\pi\to\mathcal{A}_j(\pi, \cdot):F_\varepsilon\to(L^2(\mu, \mathbb{R}), |\cdot|_w)$ is continuous for every $\varepsilon>0$, and therefore the bounded function

$$\pi\to\hat{\mathcal{A}}_j(\pi, \cdot):\mathcal{P}\to(L^2(\mu, \mathbb{R}), |\cdot|_w)$$

is ω-integrable.

Let $\mathscr{L}(\psi) := \int_{t_0}^{t_1} \psi(s)\, ds$ for all $\psi \in L^2(\mu, \mathbb{R})$. Then $\mathscr{L}$ is a bounded linear functional on $(L^2(\mu, \mathbb{R}), |\cdot|_w)$ and therefore

$$\begin{aligned}\int \omega(d\pi) \int_{t_0}^{t_1} \hat{\mathscr{A}}_j(\pi, s)\, ds &= \int \mathscr{L}(\hat{\mathscr{A}}_j(\pi, \cdot))\omega(d\pi)\\ &= \mathscr{L}\left(\int_w \hat{\mathscr{A}}_j(\pi, \cdot)\omega(d\pi)\right)\\ &= \int_{t_0}^{t_1} \left[\int_w \hat{\mathscr{A}}(\pi, \cdot)\omega(d\pi)\right](s)\, ds \quad \forall j = 1, 2, \cdots \qquad \square\end{aligned}$$

LEMMA 4.7. *Let $\mathscr{P} = \mathscr{S}_P^{\#}$, let ω and $\hat{z}$ be as in Lemma* 4.5, *and let*

$$\hat{\mathscr{A}}(r, \pi, t) := \hat{z}(\pi)(t)^T \hat{f}(t, \hat{y}_0(\pi)(t), r, \pi(t)), \qquad I(r) := \int_w \hat{\mathscr{A}}(r, \pi, \cdot)\omega(d\pi).$$

Then

$$\lim_i I(r_i)(t) = I(r)(t) \ \mu\text{-a.e.} \quad \text{if } \lim_i r_i = r \text{ in } R.$$

Proof. Let $F \subset T$ be measurable, $\alpha > 0$, and $\psi: \mathscr{S}_P^{\#} \times T \to \mathbb{R}$ such that

$$\psi(\pi, \cdot) \in L^2(\mu, \mathbb{R}), \qquad |\psi(\pi, t)| \leqq \alpha \quad \forall \pi \in \mathscr{S}_P^{\#},\ t \in F$$

and $\pi \to \psi(\pi, \cdot): \mathscr{S}_P^{\#} \to (L^2(\mu, \mathbb{R}), |\cdot|_w)$ is ω-measurable. Set

$$J(\cdot) = \int_w \psi(\pi, \cdot)\omega(d\pi).$$

Then we have

$$(4.7.1) \qquad |J(t)| \leqq \alpha\omega(\mathscr{S}_P^{\#}) \ \mu\text{-a.e.} \quad \text{in } F.$$

Indeed, assume, by way of contradiction, that there exists $H \subset F$ such that $\mu(H) > 0$ and $|J(t)| > \alpha\omega(\mathscr{S}_P^{\#})$ for all $t \in H$. Set

$$g(t) = 1/J(t) \quad \forall t \in H, \qquad g(t) = 0 \quad \forall t \in T \backslash H,$$

$$\mathscr{L}_g(\chi) = \int_{t_0}^{t_1} g(t)\chi(t)\, dt \quad \forall \chi \in L^2(\mu, \mathbb{R}).$$

Then $\mathscr{L}_g$ is a bounded linear functional on $(L^2(\mu, \mathbb{R}), |\cdot|_w)$ and therefore

$$\begin{aligned}0 < \mu(H) = \int_{t_0}^{t_1} g(t)J(t)\, dt &= \mathscr{L}_g\left(\int_w \psi(\pi, \cdot)\omega(d\pi)\right)\\ &= \int \mathscr{L}_g(\psi(\pi, \cdot))\omega(d\pi) = \int \omega(d\pi) \int_{t_0}^{t_1} g(t)\psi(\pi, t)\, dt\\ &\leqq \alpha \int \omega(d\pi) \int_{t_0}^{t_1} |g(t)|\, dt < \mu(H),\end{aligned}$$

a contradiction.

Let

$$\alpha_i(\pi)(t) := \sup \{|\hat{\mathscr{A}}(r, \pi, t) - \hat{\mathscr{A}}(r', \pi, t)| \mid \text{dist}(r, r') \leqq 1/i\}.$$

Then, for all $t \in T$, $\lim_i \alpha_i(\pi)(t) = 0$ uniformly for $\pi \in \mathcal{S}_P^\#$, and the argument used in proving Egoroff's theorem shows that for all $\varepsilon > 0$ there exist a subset A_ε of T and numbers $k(m, \varepsilon)$ such that

$$\mu(T \backslash A_\varepsilon) < \varepsilon, \qquad \alpha_i(\pi)(t) \leqq 1/m \quad \forall i \geqq k(\varepsilon, m), \quad \pi \in \mathcal{S}_P^\#, \quad t \in A_\varepsilon.$$

Thus

$$|\hat{\mathscr{A}}(r, \pi, t) - \hat{\mathscr{A}}(r', \pi, t)| \leqq 1/m \quad \forall t \in A_\varepsilon, \ \pi \in \mathcal{S}_P^\#$$

if dist $(r, r') \leqq 1/k(\varepsilon, m)$, and therefore, for each $\varepsilon > 0$, $m \in \{1, 2, \cdots\}$, and $i \geqq k(m, \varepsilon)$, the function

$$(\pi, t) \rightarrow \psi(\pi, t) := \alpha_i(\pi)(t)$$

satisfies our assumptions, with $F = A_\varepsilon$ and $\alpha = 1/m$. It follows thus, by (4.7.1), that

$$|I(r_i)(t) - I(r)(t)| = \left| \left[\int_W (\hat{\mathscr{A}}(r, \pi, \cdot) - \hat{\mathscr{A}}(r_i, \pi, \cdot)) \omega(d\pi) \right](t) \right| \leqq \frac{1}{m} \omega(\mathcal{S}_P^\#)$$

$$\mu\text{-a.e. in } A_\varepsilon$$

if dist $(r, r_i) \leqq 1/k(m, \varepsilon)$. This implies that

$$\lim_i I(r_i)(t) = I(r)(t) \ \ \mu\text{-a.e.} \quad \text{if } \lim_i r_i = r. \qquad \square$$

4.8. Proof of Theorems 2.4 and 2.5. *Step* 1. We define a subset $\mathcal{S}_\infty$ of $\mathcal{S}^\#$ in the following manner. For purposes of Theorem 2.5, we enumerate $\mathcal{S}'$ as $\{\sigma_0, \sigma_1, \cdots\}$ and set $\mathcal{S}_\infty = \text{co } S'$. For purposes of Theorem 2.4, we observe that, by Castaing's theorem [1, Thm. 5.3], there exists an at most denumerable collection $\mathcal{R}_\infty^\#$ of μ-measurable selections of $R^\#$ such that the set $\{\rho(t) \mid \rho \in \mathcal{R}_\infty^\#\}$ is dense in $R^\#(t)$ μ-almost everywhere. Let $\mathcal{I}_\infty$ be the collection of all closed subintervals of T with rational endpoints. Then the set $\mathcal{R}_\infty^\# \times \mathcal{I}_\infty$ is denumerable. For each $(\rho, I) \in \mathcal{R}_\infty^\# \times \mathcal{I}_\infty$ we define a corresponding element $\sigma \in \mathcal{S}^\#$ by

$$\sigma(t) = \delta_{\rho(t)} \quad \forall t \in I, \qquad \sigma(t) = \sigma_0(t) \quad \forall t \in T \backslash I.$$

We then adjoin to the denumerable set of such σ a denumerable dense subset of the compact metric space $\mathcal{S}^\#$ and the given point σ_0, and enumerate the entire set as $\{\sigma_0, \sigma_1, \sigma_2, \cdots\}$. We then set

$$\mathcal{S}_\infty = \text{co } \{\sigma_0, \sigma_1, \sigma_2, \cdots\}.$$

Step 2. Let $p_i : \mathbb{R}^{n+p} \rightarrow [0, \infty)$ be C^1 for $i = 1, 2, \cdots$ and such that

$$p_i(\hat{y}) = 0 \quad \text{if } |\hat{y}| \geqq 1/i \quad \text{or} \quad \hat{y} \notin \hat{V}, \qquad \int p_i(\hat{y}) \, d\hat{y} = 1,$$

where $d\hat{y}$ refers to the Lebesgue measure on $\mathbb{R}^{n+p}$. We set

$$\hat{f}_i(t, \hat{y}, r, r_P) = \int \hat{f}(t, \hat{y} - \hat{y}_1, r, r_P) p_i(\hat{y}_1) \, d\hat{y}_1.$$

We recall [6, Proof of Thm. 2.5] that the functions $\hat{f}_i(t, \cdot, \cdot, \cdot)$ and $\mathscr{D}_2 \hat{f}_i(t, \cdot, \cdot, \cdot)$ exist and are continuous, that

$$\mathscr{D}_2 \hat{f}_i(t, \hat{y}, r, r_P) \in \partial^{1/i} \hat{f}(t, \hat{y}, r, r_P) \quad \text{if } \hat{y} + \frac{1}{i} \bar{B}_{n+p} \subset \hat{V},$$

and that $\lim_i \hat{f}_i = \hat{f}$ uniformly when $\hat{y}$ is restricted to a compact set. Thus, for all sufficiently large i and each $\sigma \in \mathcal{S}_\infty$ and $\pi \in \mathcal{P}$, the differential equation

$$\hat{y}(t) = \int_{t_0}^{t} \hat{f}_i(s, \hat{y}(s), \sigma \otimes \pi(s))\, ds \quad \forall t \in T$$

has a unique solution $\hat{y}_i(\sigma \otimes \pi)$. For all such i, σ, and π, and for $\theta \in \mathcal{T}_i$, we set

$$\hat{y}_i(\sigma \otimes \pi) = (y_i(\sigma), \tilde{y}_i(\sigma \otimes \pi)), \quad \varphi(\sigma) = h(y(\sigma)(t_1)), \quad \varphi_i(\sigma) = h_i(y_i(\sigma)(t_1)),$$

$$\Phi(\sigma)(\pi) = \hat{h}(\hat{y}(\sigma \otimes \pi)(t_1)), \qquad \Phi_i(\sigma)(\pi) = \hat{h}_i(\hat{y}_i(\sigma \otimes \pi)(t_1)),$$

$$\sigma^*(\theta) = \sigma_0 + \sum_{j=1}^{i} \theta^j(\sigma_j - \sigma_0),$$

$$g_i(\theta) = \varphi_i(\sigma^*(\theta)), \qquad G_i(\theta)(\pi) = \Phi_i(\sigma^*(\theta))(\pi),$$

where (h_i), $(\hat{h}_i)$ are the sequences that define Λh, $\Lambda \hat{h}$. We easily verify (using the uniform boundedness and uniform Lipschitz continuity with respect to $\hat{v}$ of $\hat{f}$, $\hat{f}_i$, h, h_i, $\hat{h}$, $\hat{h}_i$) that

$$\lim_i y_i(\sigma) = y(\sigma), \qquad \lim_i \hat{y}_i(\sigma \otimes \pi) = \hat{y}(\sigma \otimes \pi),$$

$$\lim_i \varphi_i(\sigma) = \varphi(\sigma), \qquad \lim_i \Phi_i(\sigma)(\pi) = \Phi(\sigma)(\pi)$$

uniformly for all $\sigma \in \mathcal{S}_\infty$ and $\pi \in \mathcal{S}_P^{\#}$, and that the functions

$$\pi \to \Phi(\sigma)(\pi) : \mathcal{P} \to \mathbb{R}, \qquad \pi \to \Phi_i(\sigma)(\pi) : \mathcal{P} \to \mathbb{R}$$

are continuous (because, as previously observed, $\pi \to \sigma \otimes \pi$ and $\pi \to \hat{y}(\sigma \otimes \pi)$ are continuous for each $\sigma \in \mathcal{S}_\infty$). Thus the values of (φ, Φ) and (φ_i, Φ_i) are in $\mathbb{R}^m \times C(\mathcal{P})$. We also observe that $\hat{f}_i(t, \cdot, r, r_P)$, h_i, $\hat{h}_i$, and $\hat{y}_i$ have a common bound and Lipschitz constant that we continue to denote by c. The arguments of [4, Thms. X.3.2 and X.3.4] show that (g_i, G_i) are C^1, the functions

$$\theta \to y_i(\sigma^*(\theta)) \quad \text{and} \quad \theta \to \hat{y}_i(\sigma^*(\theta) \otimes \pi)$$

are continuous, uniformly for all θ and π, and

$$(4.8.1) \quad \begin{aligned} g_i'(\theta)e_j^i &= \int_{t_0}^{t_1} \zeta_i(\theta)(s)^T f_i(s, y_i(\sigma^*(\theta))(s), [\sigma_j - \sigma^*(\theta)](s))\, ds \\ G_i'(\theta)(\pi)e_j^i &= \int_{t_0}^{t_1} \mathfrak{z}_i(\theta)(\pi)(s)^T \hat{f}_i(s, \hat{y}_i(\sigma^*(\theta) \otimes \pi)(s), [\sigma_j - \sigma^*(\theta)] \otimes \pi(s))\, ds \end{aligned}$$

for all $\theta \in \mathcal{T}_i$ and $\pi \in \mathcal{P}$, where

$$\zeta_i(\theta) \in AC(T, \mathcal{L}(\mathbb{R}^n, \mathbb{R}^m)) \quad \text{and} \quad \mathfrak{z}_i(\theta)(\pi) \in AC(T, \mathbb{R}^{n+p})$$

are solutions of the differential equations

$$(4.8.2) \quad \begin{aligned} \zeta(t)^T &= h_i'(y_i(\sigma^*(\theta))(t_1)) + \int_t^{t_1} \zeta(s)^T \mathcal{D}_2 f_i(s, y_i(\sigma^*(\theta))(s), \sigma^*(\theta)(s))\, ds, \\ \mathfrak{z}(t)^T &= \hat{h}_i'(\hat{y}_i(\sigma^*(\theta) \otimes \pi)(t_1)) + \int_t^{t_1} \mathfrak{z}(s)^T \mathcal{D}_2 \hat{f}_i(s, \hat{y}_i(\sigma^*(\theta) \otimes \pi)(s), \sigma^*(\theta) \otimes \pi(s))\, ds. \end{aligned}$$

Furthermore, the arguments of [4, Thm. X.3.4] show that $\pi \to \mathfrak{z}_i(\theta)(\pi)$ is continuous for each i and θ.

Step 3. Let

$$\mathscr{U} = K = \mathscr{S}_\infty = \text{co}\,\{\sigma_0, \sigma_1, \sigma_2, \cdots\}, \qquad q_j = \sigma_j, \quad \delta = 1,$$
$$\mathscr{Y} = C(\mathscr{P}), \qquad C = \{g \in \mathscr{Y} \mid g(\pi) \leqq 0 \ \ \forall \pi \in \mathscr{P}\},$$
$$u_k(\theta) = \sigma^*(\theta) \quad \forall i, k = 1, 2, \cdots, \qquad \theta \in \mathscr{T}_i.$$

Then Theorem 2.6 is applicable. Alternative (a) of Theorem 2.6 yields directly alternatives (a) of Theorems 2.4 and 2.5. We will assume, therefore, that alternative (b) of Theorem 2.6 holds and will seek to derive alternatives (b) of Theorems 2.4 and 2.5.

The functionals $\ell_i^2 \in \mathscr{Y}^* = C(\mathscr{P})^*$ are represented by measures $\omega_i \in \text{frm}\,(\mathscr{P})$ and, being uniformly bounded, have a subsequence $(\ell_i^2)_{i\in\mathscr{I}}$ converging in the weak* topology to some ℓ^2 represented by some $\omega \in \text{frm}\,(\mathscr{P})$. By 2.6(b3),

$$\ell^2\Phi(\sigma_0) = \int \hat{h}(\hat{y}_0(\pi)(t_1))\omega(d\pi) = \sup\left\{\int g(\pi)\omega(d\pi) \mid g(\pi) \leqq 0 \ \ \forall \pi \in \mathscr{P}\right\}.$$

Since $\pi \to \hat{h}(\hat{y}_0(\pi)(t_1))$ is nonpositive and bounded, this implies that $\omega \in \text{frm}^+\,(\mathscr{P})$ and therefore $\ell^2\Phi(\sigma_0) = 0$ and the transversality conditions (b3) of Theorems 2.4 and 2.5 are satisfied. Furthermore, the bounded sequence $(\ell_i^1)_{i\in\mathscr{I}}$ has a subsequence $(\ell_i^1)_{i\in\mathscr{I}_1}$ converging to some $\ell^1 \in \mathbb{R}^m$. Thus the relation $0 < |\ell^1| + |\ell^2| \leqq 1$ of Theorem 2.6 yields the relation $0 < |\ell^1| + \omega(\mathscr{P}) \leqq 1$ of Theorems 2.4 and 2.5.

We next consider the terms in relation 2.6(b1). Let

$$z_i^T := \ell_i^{1T}\zeta_i(\theta_i), \quad \hat{z}_i(\pi) := \hat{\mathfrak{z}}_i(\theta_i)(\pi), \quad \sigma_i^\# := \sigma^*(\theta_i).$$

Then the sequence $(z_i)_{i\in\mathscr{I}_1}$, whose elements are uniformly bounded and have a common Lipschitz constant, has a subsequence $(z_i)_{i\in\mathscr{I}_2}$ converging uniformly to some $z \in AC(T, \mathbb{R}^n)$. Furthermore,

$$|\hat{f}_i(s, \hat{v}, (\sigma_i^\# - \sigma_0) \otimes \pi(s))| \leqq 2c \sum_{j=1}^{i} \theta_i^j \leqq 2c\gamma_i$$

for all i, s, $\hat{v}$, and π. It follows now from the first equation of (4.8.2) that z is a solution of the corresponding differential inclusions in 2.4(b1) and 2.5(b1) and, by (4.8.1),

$$\text{(4.8.3)} \qquad \lim_{i\in\mathscr{I}_2} \ell_i^1 g_i'(\theta_i)e_j^i = \int_{t_0}^{t_1} z(s)^T f(s, y_0(s), (\sigma_j - \sigma_0)(s))\, ds \quad \forall j = 1, 2, \cdots.$$

We also have, by (4.8.1) and (4.8.2),

$$\ell_i^2 G_i'(\theta_i)e_j^i = \int \omega_i(d\pi) \int_{t_0}^{t_1} \hat{z}_i(\pi)(s)^T \hat{f}_i(s, \hat{y}_i(\sigma_i^\# \otimes \pi)(s), (\sigma_j - \sigma_i^\#) \otimes \pi(s))\, ds;$$

hence

$$\text{(4.8.4)} \qquad \begin{aligned} &\lim_{i\in\mathscr{I}_2} \ell_i^2 G_i'(\theta_i)e_j^i \\ &= \lim_{i\in\mathscr{I}_2} \int \omega_i(d\pi) \int_{t_0}^{t_1} \hat{z}_i(\pi)(s)^T \hat{f}(s, \hat{y}_0(\pi)(s), (\sigma_j - \sigma_0) \otimes \pi(s))\, ds. \end{aligned}$$

Let X and $Z(\pi)$ be as defined in Lemma 4.4. We observe that the function

$$s \to \chi(s, \pi) := \hat{f}(s, \hat{y}_0(\pi)(s), (\sigma_j - \sigma_0) \otimes \pi(s))\, ds$$

is bounded, uniformly for all $\pi \in \mathcal{P}$ and $j = 1, 2, \cdots$, and it is μ-measurable. Furthermore, the function

$$\pi \to \int_{t_0}^{t_1} x(s)^T \chi(s, \pi)\, ds$$

is continuous on $\mathcal{P}$ for every $x \in X$ (since $\pi \to \sigma \otimes \pi$ is continuous for each $\sigma \in \mathcal{S}_\infty$). It follows therefore from (4.8.4) and Lemma 4.5 that there exists $\hat{z}(\pi) \in \overline{\text{co}}\, Z(\pi)$ for all $\pi \in \mathcal{P}$ such that $\pi \to \hat{z}(\pi)$ is ω-measurable and

$$\lim_{i \in \mathcal{I}_2} \ell_i^2 G_i'(\theta_i) e_j^i$$
$$= \int \omega(d\pi) \int_{t_0}^{t_1} \hat{z}(\pi)(s)^T \hat{f}(s, \hat{y}_0(\pi)(s), (\sigma_j - \sigma_0) \otimes \pi(s))\, ds \quad \forall j = 1, 2, \cdots.$$

This last statement and (4.8.3), together with relation 2.6(b1), now yield 2.4(b1), 2.5(b1), and the inequality

$$\int_{t_0}^{t_1} z(s)^T f(s, y_0(s), (\sigma_j - \sigma_0)(s))\, ds$$
$$+ \int \omega(d\pi) \int_{t_0}^{t_1} \hat{z}(\pi)(s)^T \hat{f}(s, \hat{y}_0(\pi)(s), (\sigma_j - \sigma_0) \otimes \pi(s))\, ds \geqq 0 \tag{4.8.5}$$
$$\forall j = 1, 2, \cdots,$$

which yields, in particular, statement 2.5(b2.1).

Step 4. Let

$$\mathcal{A}(r, t) := z(t)^T f(t, y_0(t), r), \qquad \mathcal{A}_j(t) = \mathcal{A}((\sigma_j - \sigma_0)(t), t),$$

$\hat{\mathcal{A}}_j(\pi, t)$ be as defined in Lemma 4.6, and

$$\mathcal{B}_j := \int_w \hat{\mathcal{A}}_j(\pi, \cdot)\, \omega(d\pi).$$

By (4.8.5) and Lemma 4.6, we have

$$\int_{t_0}^{t_1} [\mathcal{A}_j(s) + B_j(s)]\, ds \geqq 0 \quad \forall j = 1, 2, \cdots.$$

Let L be the set of the common Lebesgue points of $\mathcal{A}_j + \mathcal{B}_j$ for all j that is of full μ-measure, let $t \in L \cap (t_0, t_1)$, and let (a_k) and (b_k) be sequences of rational numbers in T such that

$$a_k < t < b_k, \qquad \lim_k (b_k - a_k) = 0.$$

Then, for all $j = 1, 2, \cdots$,

$$\lim_k \frac{1}{b_k - a_k} \int_{a_k}^{b_k} [\mathcal{A}_j(s) + \mathcal{B}_j(s)]\, ds = \mathcal{A}_j(t) + B_j(t). \tag{4.8.6}$$

We observe that, for purposes of Theorem 2.4, for each $\rho \in \mathcal{R}_\infty^\#$ and k there exists some j such that

$$\sigma_j(s) = \delta_{\rho(s)} \ \text{ if } a_k \leqq s \leqq b_k, \qquad \sigma_j(s) = \sigma_0(s) \ \text{ if } s \in T \backslash [a_k, b_k].$$

For purposes of Theorem 2.5, and under the assumption of 2.5(b2.2), for each $\sigma' \in \mathscr{S}'$ and k there exists some j such that

$$\sigma_j(s) = \sigma'(s) \quad \text{if } a_k \leqq s \leqq b_k, \qquad \sigma_j(s) = \sigma_0(s) \quad \text{if } s \in T \backslash [a_k, b_k].$$

Since $\mathscr{A}_j(s) = \mathscr{B}_j(s) = 0$ for $s \in T\backslash[a_k, b_k]$, it follows from (4.8.5) and (4.8.6) that, for all $\sigma \in \mathscr{R}^\#_\infty$, respectively, $\sigma \in \mathscr{S}'$ and all $t \in L$, we have

$$\begin{aligned} &z(t)^T f(t, y_0(t), \sigma(t)) + \left[\int_W \hat{z}(\pi)(\cdot)^T \hat{f}(\cdot, \hat{y}_0(\pi)(\cdot), \sigma \otimes \pi(\cdot))\omega(d\pi)\right](t) \\ &\qquad \geqq z(t)^T f(t, y_0(t), \sigma_0(t)) + \left[\int_W \hat{z}(\pi)(\cdot)^T \hat{f}(\cdot, \hat{y}(\pi)(\cdot), \sigma_0 \otimes \pi(\cdot))\omega(d\pi)\right](t). \end{aligned} \tag{4.8.7}$$

Within the context of Theorem 2.5, this proves statement 2.5(b2.2) and completes the proof of Theorem 2.5.

Step 5. We henceforth consider statement (b2) of Theorem 2.4 and set $\mathscr{P} = \mathscr{S}^\#_P$. By choice, the set $\{\sigma_1, \sigma_2, \cdots\}$ is dense in $\mathscr{S}^\#$, and the function $\sigma \to \sigma \otimes \pi : \mathscr{S}^\# \to \hat{\mathscr{S}}^\#$ is continuous for each $\pi \in \mathscr{S}^\#_P$. Thus both integrals over T that appear in (4.8.5) are continuous functions of σ over $\mathscr{S}^\#$ if σ_j is replaced by σ. Therefore, inequality (4.8.5) remains valid when σ_j is replaced by any $\sigma \in \mathscr{S}^\#$. This proves relation 2.4(b2.1). It follows from (4.8.7) and Lemma 4.7 that, if $R^\#(t) = R$ for all $t \in T$ then, for all $r \in R$,

$$\begin{aligned} &z(t)^T f(t, y_0(t), r) + \left[\int_W \hat{z}(\pi)(\cdot)^T \hat{f}(\cdot, \hat{y}_0(\pi)(\cdot), r, \pi(\cdot))\omega(d\pi)\right](t) \\ &\qquad \geqq z(t)^T f(t, y_0(t), \sigma_0(t)) + \left[\int_W \hat{z}(\pi)(\cdot)^T \hat{f}(\cdot, \hat{y}(\pi)(\cdot), \sigma_0(\cdot), \pi(\cdot))\omega(d\pi)\right](t) \\ &\qquad\qquad \mu\text{-a.e.} \end{aligned}$$

This proves statement 2.4(b2.2).

We next derive the pointwise maximum principle (b2.3(i)) of Theorem 2.4. As before, let

$$\mathscr{A}(r, t) := z(t)^T f(t, y_0(t), r), \qquad \hat{\mathscr{A}}(r, \pi, t) := \hat{z}(\pi)(t)^T \hat{f}(t, \hat{y}_0(\pi)(t), r, \pi(t)).$$

For each $\rho \in \mathscr{R}^\#_\infty$ and $\pi \in \operatorname{supp}(\omega)$, the functions

$$t \to \mathscr{A}(\delta_{\rho(t)} - \sigma_0(t), t) \quad \text{and} \quad t \to \hat{\mathscr{A}}(\delta_{\rho(t)} - \sigma_0(t), \pi, t)$$

are μ-integrable and, therefore, the set of their common Lebesgue points is of full μ-measure. Since $\mathscr{R}^\#_\infty$ and (by 2.4(b3)) $\operatorname{supp}(\omega)$ are at most denumerable, the set L of their common Lebesgue points for all $\rho \in \mathscr{R}^\#_\infty$ and $\pi \in \operatorname{supp}(\omega)$ is also of full μ-measure. Now let $t \in L \cap (t_0, t_1)$, and let (a_k) and (b_k) be sequences of rational numbers in T such that

$$a_k < t < b_k, \qquad \lim_k (b_k - a_k) = 0.$$

Then, for all $\sigma \in \{\sigma_0, \sigma_1, \cdots\}$ and $\pi \in \operatorname{supp}(\omega)$,

$$\lim_k \frac{1}{b_k - a_k} \int_{a_k}^{b_k} \mathscr{A}(\sigma(s), s)\, ds = \mathscr{A}(\sigma(t), t)$$

$$\lim_k \frac{1}{b_k - a_k} \int_{a_k}^{b_k} \hat{\mathscr{A}}(\sigma(s), \pi, s)\, ds = \hat{\mathscr{A}}(\sigma(t), \pi, t).$$

We observe that for each $\rho \in \mathcal{R}^{\#}_{\infty}$ and k there exists some j such that

$$\sigma_j(s) = \delta_{\rho(s)} \quad \text{if } a_k \leqq s \leqq b_k, \qquad \sigma_j(s) = \sigma_0(s) \quad \text{if } s \in T \backslash [a_k, b_k].$$

It follows therefore from (4.8.5) that, for all $\rho \in \mathcal{R}^{\#}_{\infty}$ and $t \in L$,

$$(4.8.8) \quad \mathcal{A}(\rho(t), t) + \int \hat{\mathcal{A}}(\rho(t), \pi, t)\omega(d\pi) \geqq \mathcal{A}(\sigma_0(t), t) + \int \hat{\mathcal{A}}(\sigma_0(t), \pi, t)\omega(d\pi).$$

The functions $f(t, \mathbf{y}, \cdot)$ and $\hat{f}(t, \hat{\mathbf{y}}, \cdot, r_P)$ are equicontinuous and $\pi(t)$ are probability measures. Thus, for each $t \in L$, the functions

$$r \to \mathcal{A}(r, t) \quad \text{and} \quad r \to \int \hat{\mathcal{A}}(r, \pi, t)\omega(d\pi)$$

are continuous. Since the set $\{\rho(t) \mid \rho \in \mathcal{R}^{\#}_{\infty}\}$ is dense in $R^{\#}(t)$ μ-almost everywhere and $\sigma_0(t)$ is supported on $R^{\#}(t)$, we conclude from (4.8.8) that, for μ-almost all t and for all $r \in R^{\#}(t)$,

$$\mathcal{A}(\sigma_0(t), t) + \int \hat{\mathcal{A}}(\sigma_0(t), \pi, t)\omega(d\pi) \leqq \mathcal{A}(r, t) + \int \hat{\mathcal{A}}(r, \pi, t)\omega(d\pi).$$

This proves statement (b2.3(i)) of Theorem 2.4.

We finally derive the maximum principle (b2.3(ii)) of Theorem 2.4. Since $\hat{f}(t, \hat{\mathbf{y}}, r, r_P) = \not{f}_1(t, \hat{\mathbf{y}}, r) + \not{f}_2(t, \hat{\mathbf{y}}, r_P)$, the maximum principle (b2.1) yields

$$\int_{t_0}^{t_1} z(s)^T f(s, y_0(s), (\sigma - \sigma_0)(s))\, ds$$
$$+ \int \omega(d\pi) \int_{t_0}^{t_1} \hat{z}(\pi)(s)^T \hat{\not{f}}_1(s, \hat{\mathbf{y}}_0(\pi)(s), (\sigma - \sigma_0)(s))\, ds \geqq 0 \quad \forall \sigma \in \mathcal{S}^{\#}.$$

The function $(s, \pi) \to \hat{z}(\pi)(s)$, ω-measurable in π and continuous in s, and the function $(s, \pi) \to \hat{\not{f}}_1(s, \hat{y}_0(\pi)(s), (\sigma - \sigma_0)(s))$, μ-measurable in s and continuous in π, are both $\mu \times \omega$-measurable. We may therefore apply Fubini's theorem to the double integral above and obtain

$$(4.8.9) \qquad \int_{t_0}^{t_1} E(s, (\sigma - \sigma_0)(s))\, ds \geqq 0 \quad \forall \sigma \in \mathcal{S}^{\#},$$

where

$$E(s, (\sigma - \sigma_0)(s))$$
$$= z(s)^T f(s, y_0(s), (\sigma - \sigma_0)(s)) + \int \hat{z}(\pi)(s)^T \hat{\not{f}}_1(s, \hat{y}_0(\pi)(s), (\sigma - \sigma_0)(s))\omega(d\pi).$$

If we choose $\rho \in \mathcal{R}^{\#}_{\infty}$ then, for any μ-measurable set F and

$$\sigma(s) = \delta_{\rho(s)} \quad \forall s \in F, \qquad \sigma(s) = \sigma_0(s) \quad \forall s \in T \backslash F,$$

relation (4.8.9) yields $\int_F [E(s, \rho(s)) - E(s, \sigma_0(s))\, ds \geqq 0$, hence

$$E(t, \rho(t)) - E(s, \sigma_0(t)) \geqq 0 \quad \mu\text{-a.e.}$$

Since $\mathcal{R}^{\#}_{\infty}$ is denumerable and $\{\rho(t) \mid \rho \in \mathcal{R}^{\#}_{\infty}\}$ dense in $R^{\#}(t)$ μ-almost everywhere, we conclude that the maximum principle 2.4(b2.3(ii)) is valid. This concludes the proof of Theorem 2.4. □

REFERENCES

[1] CH. CASTAING, *Sur les multi-applications mesurables*, Revue Française d'Informatique et de Recherche Opérationnelle, No. 1, 1967.

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

[3] E. PARZEN, *Stochastic Processes*, Holden-Day, San Francisco, CA, 1962.

[4] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[5] ———, *Necessary conditions without differentiability assumptions in unilateral control problems*, J. Differential Equations, 21 (1976), pp. 25-38.

[6] ———, *Derivate containers, inverse functions, and controllability*, in Calculus of Variations and Control Theory, D. L. Russell, ed., Academic Press, New York, 1976.

[7] ———, *Optimization and controllability without differentiability assumptions*, SIAM J. Control Optim., 21 (1983), pp. 837-855.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 3, pp. 702-730, May 1991

MARTINGALE AND DUALITY METHODS FOR UTILITY MAXIMIZATION IN AN INCOMPLETE MARKET*

IOANNIS KARATZAS†, JOHN P. LEHOCZKY‡, STEVEN E. SHREVE§, AND GAN-LIN XU¶

Abstract. The problem of maximizing the expected utility from terminal wealth is well understood in the context of a complete financial market. This paper studies the same problem in an *incomplete market* containing a bond and a finite number of stocks whose prices are driven by a multidimensional Brownian motion process W. The coefficients of the bond and stock processes are adapted to the filtration (history) of W, and incompleteness arises when the number of stocks is *strictly smaller* than the dimension of W. It is shown that there is a way to complete the market by introducing additional "fictitious" stocks so that the optimal portfolio for the thus completed market *coincides* with the optimal portfolio for the original incomplete market. The notion of a "least favorable" completion is introduced and is shown to be closely related to the existence question for an optimal portfolio in the incomplete market. This notion is expounded upon using martingale techniques; several equivalent characterizations are provided for it, examples are studied in detail, and a fairly general existence result for an optimal portfolio is established based on convex duality theory.

Key words. incomplete markets, portfolio processes, stochastic control, convex duality, utility maximization

AMS(MOS) subject classifications. primary 93E20; secondary 60G44, 90A16, 49B60

1. Introduction.

This paper studies the problem of an agent who receives a deterministic initial capital, which he must then invest in an incomplete market so as to maximize the expected utility of his wealth at a prespecified final time. The market consists of a bond and m stocks, the latter being driven by a d-dimensional Brownian motion. In such a model, *incompleteness arises when m is strictly smaller than d.* The market coefficients, i.e., the interest rate, the rates of stock appreciation, and the stock volatility coefficients, are random processes adapted to the full d-dimensional Brownian motion. When $m < d$, it is typically not possible to construct a portfolio consisting of the bond and the m available stocks, so as to completely hedge the risk associated with these coefficient processes.

Our model is not Markov, and so the Bellman equation of dynamic programming is inadequate for its analysis. Using the Bellman equation, Svensson [21] has treated an infinite-horizon, incomplete, Markov model with an income process. He derives first-order conditions and obtains explicit solutions when the utility for consumption has constant absolute risk aversion. Duffie and Jackson [4] provide a similar analysis of a finite-horizon, incomplete, Markov model.

* Received by the editors October 25, 1989; accepted for publication (in revised form) July 5, 1990. An abridged version of this paper was presented at the *Workshop on Mathematical Theory of Modern Financial Markets*, Mathematical Sciences Institute, Cornell University, July 1989, and at the *Ninth International Conference "Analysis and Optimization of Systems,"* Antibes, France, June 1990.

† Department of Mathematics, Rutgers University, New Bruswick, New Jersey 08903. The author is on leave from Columbia University. This research was supported by National Science Foundation grant DMS-87-23078.

‡ Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. This research was supported by National Science Foundation grant DMS-87-02537.

§ Department of Mathematics, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. This research was supported by National Science Foundation grant DMS-87-02537.

¶ Global Portfolio Research Department, Daiwa Securities Trust Company, Jersey City, New Jersey 07302.

The principal result of this paper, Theorem 12.5, provides conditions under which an optimal portfolio exists in an incomplete market, and characterizes this optimal portfolio in terms of the solution to a dual optimization problem.

In § § 2–5, we define the utility maximization problem faced by the agent. In § 6 we present the solution when the market is complete $(m = d)$, and complete hedging is possible. This solution proceeds in three steps. First, on the underlying probability space we determine a new measure which discounts the growth inherent in the market; under this measure, the expected value of the final wealth attained by any reasonable portfolio is equal to the initial endowment. Second, among all random variables whose expectation under the new measure is equal to the initial endowment, a most desirable one is determined. Third, it is shown that a portfolio can be constructed that attains this most desirable random variable as its terminal wealth; this portfolio is optimal. A *complete market* is one in which the agent can construct a portfolio that attains as final wealth any random variable with expectation under the new measure equal to the initial endowment. Because such a construction is possible, it is said that the agent can *hedge against the risk* associated with this market. Mathematically, the construction of a portfolio uses the fact that any martingale with respect to a Brownian filtration can be represented as a stochastic integral with respect to the Brownian motion; the integrand in this representation leads to the portfolio we are seeking. However, if there are fewer than d stocks, this line of argument fails.

In § 7 we introduce a convenient way of thinking about an incomplete market: *fictitious completion.* When there are fewer than d stocks, then we augment the stocks with certain fictitious ones so as to create a complete market. If the fictitious stocks have a high appreciation rate, then under an optimal portfolio the agent will hold a long position in them, but if they have a low (even negative) appreciation rate, then he will hold a short position. Thus we would expect to be able to adjust the appreciation rates of the fictitious stocks so that the agent, by optimal choice, does not invest in them at all. These judiciously chosen fictitious stocks allow us to write down the complete market solution for the utility maximization problem but are superfluous in the actual implementation of the optimal portfolio, which must then also be optimal for the original incomplete market. The fictitious completion with the above property is the least advantageous to the agent, because the portfolio which is optimal under this completion is available to him under every other fictitious completion. We thus have the notion of a *least favorable fictitious completion*: for every fictitious completion we compute the portfolio which maximizes the expected utility of final wealth, and then we choose the completion which makes this maximum expected utility as small as possible.

As explained in § 7, a convenient way to parametrize fictitious completions of an incomplete market is by a certain space of continuous local martingales, each local martingale being the Radon–Nikodym derivative process of the new measure alluded to in the earlier discussion of complete markets. This kind of parametrization is studied in § 8, and several pertinent results are established. It is also desirable to characterize the local martingale corresponding to the least favorable fictitious completion, and to show that it gives rise to an optimal portfolio in the original incomplete problem; this program is carried out in § 9, where various such equivalent characterizations are provided. Section 10 studies two examples in which the least favorable fictitious completion can be computed fairly explicitly. In the first example the utility function is logarithmic, and it is discovered that the fictitious stocks in the least favorable completion should have rates of appreciation equal to the interest rate of the bond. This is a very general result, insensitive to the nature of the dependence of market coefficients on the driving Brownian motion. In the second example it is assumed that the utility function is of the power form, and that the driving Brownian motion splits into two independent parts; the first part drives the stock processes, whose coefficients are adapted solely to the second part. The least favorable local martingale is exhibited as the solution to a martingale representation problem, and the optimal portfolio is found to be given by the formula already known to be correct for deterministic model coefficients.

In § 11 we introduce an auxiliary optimization problem involving the family of local martingales which characterize fictitious completions; this problem is "dual" to the "primal" utility maximization question of § 5, in the sense of convex duality. We study the relation between the primal and dual problems, and explain how a solution to the latter induces one for the former. The question of existence in the dual problem is tantamount to the existence of a least favorable fictitious completion; it is dealt with in § 12, by the use of methods from convex analysis.

Our model for the financial market can be traced back to Merton [16], [17] and Samuelson [20]. The modern mathematical approach to portfolio management in complete markets, built around the ideas of equivalent martingale measures and the creation of portfolios from martingale representation theorems, began with Harrison and Kreps [6] and was further developed by Harrison and Pliska [7], [8], in the context of option pricing. Pliska [19], Cox and Huang [2], [3], and Karatzas, Lehoczky, and Shreve [13] adapted the martingale ideas to problems of utility maximization. Much of this development appears in § 5.8 of Karatzas and Shreve [14], from which § 6 of the present paper is drawn (see also the review article of Karatzas [12] for a survey of financial economics problems in complete markets). An extension of the papers above to infinite horizon problems is reported by Huang and Pagès [11].

A first step toward a martingale analysis of incomplete markets was taken by Pagès [18], who considered a Brownian model in which the number of stocks was strictly less than the dimension of the driving Brownian motion. However, the coefficients of the bond and stock prices in this model were allowed to depend on the underlying Brownian model only through the bond and stock prices themselves. Thus, the vector of bond and stock prices formed a Markov process. This specialization created an essentially complete market, and thus it avoided the more interesting case of a market with genuinely unhedgeable risk. However, Pagès did characterize the class of equivalent martingale measures which could arise in an incomplete model, and this laid the groundwork for further developments (e.g., Lemma 8.2 in this paper). A more substantial step was taken by He and Pearson [9] in a discrete-time, finite probability space model, where the authors proposed finding the optimal intermediate consumption and terminal wealth corresponding to each of the equivalent martingale measures, and then searching over those policies to find a pair yielding the minimum expected total utility. Using separating hyperplane arguments, they were able to show that the total utility obtained by this two-step "minimax" process is the optimal total value for the incomplete problem.

He and Pearson have also studied the incomplete problem in a continuous-time, Brownian model. In an early version of He and Pearson [10], the authors consider Pagès's characterization of the family of equivalent martingale measures and search over this family for a "minimax" equivalent martingale measure, which would lead them to the optimal consumption and portfolio processes just as in a complete market. The martingale associated with this measure would create the "Arrow–Debreu" state prices in the incomplete model. However, the continuous-time model is more subtle than one might expect, and although it is now clear that Arrow–Debreu state prices exist for the incomplete model under some assumptions, it is not clear that they are associated with a martingale.

The present paper uses local martingales rather than martingales to address the issue of market incompleteness in continuous-time models. This work was motivated by the aforementioned previous version of He and Pearson [10], and by the use of local martingale methods introduced by Xu [22] in the study of incompleteness induced by a prohibition on the short-selling of stocks. Using the stochastic duality theory of Bismut [1], Xu formulated a dual problem whose solution could be shown to exist and could then be used to obtain existence and characterization of the solution of the original problem. As this paper shows, such duality methods can also be used in the traditional incomplete Brownian market model. While we still do not know if the minimax equivalent martingale measure sought by He and Pearson exists in any generality (see, however, the note following the references), we show here that the solution to Bismut's dual problem is a "least favorable local martingale" which can be used to generate a sequence of equivalent measures. The existence of this least favorable local martingale is sufficient for the study of many models. A notable exception is the incomplete market model in which the agent's endowment is a stochastic process; we do not know how to obtain the existence and a characterization of the optimal policy for such a model in terms of a least favorable local martingale, unless it is actually a martingale.

He and Pearson [10] have incorporated Xu's local martingale techniques into their original work. They report the existence of an optimal portfolio for the terminal wealth utility maximization problem when the index of relative risk aversion is everywhere greater than or equal to one, and they report similar results for the problem with intermediate consumption and consumption at the terminal time when the index of relative risk aversion is everywhere less than or equal to one. Our paper deals only with the case of terminal wealth utility maximization when the index of relative risk aversion is everywhere less than or equal to one; the generalization to also allow for intermediate consumption is straightforward. Whereas He and Pearson [10] assume that some augmentation of the market model will result in Markov prices, we allow general Itô price processes. He and Pearson [10] do not provide the full set of equivalent conditions contained in our Theorem 9.4, nor do they use our assumption (4.8). This assumption and the introduction of the set $K_1(\sigma)$ play a fundamental role in our proof of Theorem 9.4.

Remark on Notation: We denote by "standing assumption" those conditions that are always in force; they will not be cited in the theorems. The Standing Assumptions are 2.1, 2.2, 2.3, 4.1, and 5.1. We denote by "assumption" those conditions which are in force only when theorems specifically cite them.

2. The market model. We adopt a model for the financial market consisting of one bond with price $P_0(t)$ given by

$$dP_0(t) = r(t)P_0(t)\,dt, \qquad P_0(0) = 1, \tag{2.1}$$

and m stocks with prices per share $P_i(t)$, $i = 1, \cdots, m$, satisfying the equations

$$dP_i(t) = P_i(t)\left[b_i(t)\,dt + \sum_{j=1}^{d} \sigma_{ij}(t)\,dW_j(t)\right], \qquad i = 1, \cdots, m. \tag{2.2}$$

Here $W = (W_1, \cdots, W_d)^*$ is a d-dimensional Brownian motion on a probability space $(\Omega, \mathscr{F}, P)$, and we denote by $\{\mathscr{F}_t\}$ the P-augmentation of the filtration generated by W. It is assumed throughout that $d \geqq m$, i.e., the number of sources of uncertainty in the model is at least as large as the number of stocks available for investment.

The *interest rate* $r(t)$, the vector $b(t)=(b_1(t),\cdots,b_m(t))^*$ of *stock appreciation rates*, and the *volatility matrix* $\sigma(t)=\{\sigma_{ij}(t)\}_{1\leqq i\leqq m,\,1\leqq j\leqq d}$ are the *coefficients of the model.* They are taken to be progressively measurable with respect to $\{\mathcal{F}_t\}$.

Standing Assumption 2.1. *$\int_0^T \|b(t)\|\,dt<\infty$, $\int_0^T |r(t)|\,dt\leqq L$ hold almost surely, for some given real constant $L>0$.*

The positive constant T in Standing Assumption 2.1 is the *terminal time* for the problem. All processes are defined on $[0, T]$.

Standing Assumption 2.2. *The matrix $\sigma(t)$ has full rank for every t.*

As a result of Standing Assumption 2.2, the matrix $(\sigma(t)\sigma^*(t))^{-1}$ and the *relative risk* process

$$\theta(t)\triangleq\sigma^*(t)(\sigma(t)\sigma^*(t))^{-1}[b(t)-r(t)\mathbf{1}_m] \tag{2.3}$$

are defined. Throughout this paper, we denote by $\mathbf{1}_k$ the k-dimensional vector whose every component is one. In addition to Standing Assumptions 2.1 and 2.2, the following assumption will be made throughout.

Standing Assumption 2.3. *$\int_0^T \|\theta(t)\|^2\,dt<\infty$, almost surely P.*

We shall have occasion to use the so-called *discount process*

$$\beta(t)\triangleq\frac{1}{P_0(t)}=\exp\left\{-\int_0^t r(s)\,ds\right\}, \tag{2.4}$$

as well as the process

$$W_0(t)\triangleq W(t)+\int_0^t \theta(s)\,ds \tag{2.5}$$

and the exponential local martingale

$$Z_0(t)\triangleq\exp\left\{-\int_0^t \theta^*(s)\,dW(s)-\frac{1}{2}\int_0^t \|\theta(s)\|^2\,ds\right\}. \tag{2.6}$$

DEFINITION 2.4. A financial market as above will be called *complete* if $m=d$, and *incomplete* if $m<d$.

3. Portfolio and wealth processes. A *portfolio process* $\pi(t)=(\pi_1(t),\cdots,\pi_m(t))^*$ is an $\mathbb{R}^m$-valued, $\{\mathcal{F}_t\}$-adapted process satisfying

$$\int_0^T \|\sigma^*(t)\pi(t)\|^2\,dt<\infty \quad \text{a.s. } P. \tag{3.1}$$

We regard $\pi_i(t)$ as the proportion of an agent's wealth invested in stock i at time t; the remaining proportion $1-\pi^*(t)\mathbf{1}_m=1-\sum_{i=1}^m \pi_i(t)$ is invested in the bond. We do *not* constrain these proportions to take values in the interval $[0, 1]$; in other words, we allow both short-selling of stocks, and borrowing at the interest rate of the bond. For a given, nonrandom, initial wealth $x>0$, let $X^{x,\pi}$ denote the *wealth process* corresponding to a portfolio π defined by $X^{x,\pi}(0)=x$ and

$$\begin{aligned} dX^{x,\pi}(t)&=r(t)X^{x,\pi}(t)\,dt+X^{x,\pi}(t)\pi^*(t)[(b(t)-r(t)\mathbf{1}_m)\,dt+\sigma(t)dW(t)] \\ &=r(t)X^{x,\pi}(t)\,dt+X^{x,\pi}(t)\pi^*(t)\sigma(t)dW_0(t). \end{aligned} \tag{3.2}$$

In other words,

$$\begin{aligned} \beta(t)X^{x,\pi}(t)&=x\exp\left\{\int_0^t \pi^*(s)\sigma(s)\,dW_0(s)-\frac{1}{2}\int_0^t \|\sigma^*(s)\pi(s)\|^2\,ds\right\} \\ &=x+\int_0^t \beta(s)X^{x,\pi}(s)\pi^*(s)\sigma(s)\,dW_0(s), \qquad 0\leqq t\leqq T. \end{aligned} \tag{3.3}$$

Remark 3.1. An application of Itô's rule to the product of the processes Z_0 and $\beta X^{x,\pi}$ of (2.6), (3.3) leads to

$$\beta(t)Z_0(t)X^{x,\pi}(t)=x+\int_0^t \beta(s)Z_0(s)X^{x,\pi}(s)(\sigma^*(s)\pi(s)-\theta(s))^*\,dW(s). \tag{3.4}$$

This shows, in particular, that the process $\beta Z_0 X^{x,\pi}$ is a nonnegative local martingale, hence a supermartingale.

4. Utility functions. The agent in our model has a utility function $U:(0,\infty)\to\mathbb{R}$ for wealth. We make the following assumption throughout.

Standing Assumption 4.1. *U is strictly increasing, strictly concave, continuous and continuously differentiable, and satisfies*

$$U'(0)\triangleq\lim_{x\downarrow 0}U'(x)=\infty,\qquad U'(\infty)\triangleq\lim_{x\to\infty}U'(x)=0. \tag{4.1}$$

The (continuous, strictly decreasing) inverse of the function U' will be denoted by $I:(0,\infty)\to(0,\infty)$; by analogy with (4.1), it satisfies

$$I(0)\triangleq\lim_{y\downarrow 0}I(y)=\infty,\qquad I(\infty)\triangleq\lim_{y\to\infty}I(y)=0. \tag{4.2}$$

We introduce also the function

$$\tilde U(y)\triangleq\max_{x>0}[U(x)-xy]=U(I(y))-yI(y),\qquad 0<y<\infty, \tag{4.3}$$

which is the Legendre transform of $-U(-x)$, with U extended to be $-\infty$ on the negative real axis. The function $\tilde U$ is strictly decreasing, strictly convex, and satisfies

$$\tilde U'(y)=-I(y),\qquad 0<y<\infty, \tag{4.4}$$

$$U(x)=\min_{y>0}[\tilde U(y)+xy]=\tilde U(U'(x))+xU'(x),\qquad 0<x<\infty. \tag{4.5}$$

The useful inequalities

$$U(I(y))\geqq U(x)+y[I(y)-x]\qquad \forall x>0, y>0, \tag{4.6}$$

$$\tilde U(U'(x))\leqq \tilde U(y)-x[U'(x)-y]\quad \forall x>0, y>0 \tag{4.7}$$

then follow directly from (4.3), (4.5).

The monotonicity of U and $\tilde U$ guarantees that the limits

$$U(0)\triangleq\lim_{x\downarrow 0}U(x),\qquad U(\infty)\triangleq\lim_{x\to\infty}U(x),$$

$$\tilde U(0)\triangleq\lim_{y\downarrow 0}\tilde U(y),\qquad \tilde U(\infty)\triangleq\lim_{y\to\infty}\tilde U(y)$$

exist in the extended real-number system.

LEMMA 4.2. $U(0)=\tilde U(\infty)$, $\tilde U(0)=U(\infty)$.

Proof. It follows from (4.3) that $\tilde U(\infty)\leqq\lim_{y\to\infty}U(I(y))=U(0)$, as well as

$$\tilde U(\infty)\geqq\lim_{z\to\infty}\left[U\left(\frac{\varepsilon}{z}\right)-\varepsilon\right]=U(0)-\varepsilon\quad \forall\varepsilon>0,$$

whence $\tilde U(\infty)=U(0)$. Similarly, it follows from (4.5) that $U(\infty)\geqq\lim_{x\to\infty}\tilde U(U'(x))=\tilde U(0)$, as well as

$$U(\infty)\leqq\lim_{\xi\to\infty}\left[\tilde U\left(\frac{\varepsilon}{\xi}\right)+\varepsilon\right]=\tilde U(0)+\varepsilon\quad \forall\varepsilon>0,$$

whence $U(\infty)=\tilde U(0)$. □

We will sometimes impose the following condition on the utility function U.

Assumption 4.3. For some $\alpha \in (0,1)$, $\gamma \in (1,\infty)$, we have

$$\alpha U'(x) \geqq U'(\gamma x) \quad \forall x \in (0,\infty). \tag{4.8}$$

Quite obviously, Assumption 4.3 is satisfied by the utility functions $U(x) = \log x$ and $U(x) = x^\delta/\delta$, with $\delta < 1$, $\delta \neq 0$. Upon replacing x by $I(y)$ in (4.8), and then applying I to both sides of the resulting inequality, we see that Assumption 4.3 is equivalent to the condition

$$I(\alpha y) \leqq \gamma I(y) \quad \forall y \in (0,\infty), \tag*{(4.8)$'$}$$

for some $\alpha \in (0,1)$, $\gamma \in (1,\infty)$. By iterating (4.8)$'$, we obtain the apparently stronger statement

$$\forall \alpha \in (0,1), \quad \exists \gamma \in (1,\infty), \quad \text{such that} \quad I(\alpha y) \leqq \gamma I(y) \quad \forall y \in (0,\infty). \tag{4.9}$$

5. The utility maximization problem. For a given utility function U and a given initial capital $x > 0$, the stochastic control problem considered in this paper is the following: *to maximize the expected utility from terminal wealth* $EU(X^{x,\pi}(T))$, over the class $\mathscr{A}(x)$ of portfolio processes π that satisfy

$$E(U(X^{x,\pi}(T)))^- < \infty, \tag{5.1}$$

where $a^- \triangleq \max\{-a, 0\}$. The value function of this problem is denoted by

$$V(x) \triangleq \sup_{\pi \in \mathscr{A}(x)} EU(X^{x,\pi}(T)). \tag{5.2}$$

To be sure this problem is meaningful, we make the following assumption throughout.

Standing Assumption 5.1. $V(x) < \infty$, *for all* $x \in (0,\infty)$.

A portfolio process $\pi \in \mathscr{A}(x)$ which attains the supremum in (5.2) is called *optimal*. In §§ 9, 11, and 12 we provide conditions that ensure the existence of optimal portfolios, as well as various characterizations of optimality. Some examples, in which optimal portfolios can be computed explicitly, appear in § 10.

Remark 5.2. In the case of a market model for which the relative risk process $\theta(\cdot)$ of (2.4) satisfies the condition

$$\int_0^T \|\theta(t)\|^2 \, dt \leqq C \quad \text{a.s.} \tag{5.3}$$

for some given real constant $C > 0$, a sufficient condition for Standing Assumption 5.1 is

$$U(x) \leqq k_1 + k_2 x^\delta \quad \forall x \in (0,\infty) \tag{5.4}$$

for some $k_1 > 0$, $k_2 > 0$, $\delta \in (0,1)$.

Indeed, the process Z_0 of (2.6) is then a martingale, and $W_0(\cdot)$ is a Brownian motion under the probability measure $P_0(A) = E[Z_0(T)1_A]$ on $\mathscr{F}_T$ (the Girsanov theorem; cf. Karatzas and Shreve [14, § 3.5]). For any $p \in (1, 1/\delta]$ and suitable constants $c_1 > 0$, $c_2 > 0$, we have

$$U_+^p(x) \leqq c_1 + c_2 x^{\delta p} \quad \forall x \in (0,\infty) \tag{5.5}$$

from (5.4), where $U_+(x) \triangleq \max\{U(x), 0\}$. Also, we have

$$\begin{aligned}(X^{x,\pi}(T))^{\delta p} &= x^{\delta p} \cdot \exp\left[\delta p \int_0^T r(s)\,ds - \frac{\delta p(1-\delta p)}{2}\int_0^T \|\sigma^*(s)\pi(s)\|^2\,ds\right] \\ &\quad \cdot \exp\left[\delta p \int_0^T \pi^*(s)\sigma(s)\,dW_0(s) - \frac{\delta^2 p^2}{2}\int_0^T \|\sigma^*(s)\pi(s)\|^2\,ds\right] \\ &\leqq (e^L x)^{\delta p} \cdot \exp\left[\delta p \int_0^T \pi^*(s)\sigma(s)\,dW_0(s)\right. \\ &\quad \left. -\frac{1}{2}\delta^2 p^2 \int_0^T \|\sigma^*(s)\pi(s)\|^2\,ds\right]\end{aligned} \tag{5.6}$$

from (3.3), and

$$\begin{aligned}E_0 Z^{-q}(t) &= E_0\left[\exp\left\{q\int_0^T \theta^*(s)\,dW_0(s) - \frac{q^2}{2}\int_0^T \|\theta(s)\|^2\,ds\right\}\right. \\ &\qquad\qquad \left. \cdot \exp\left\{\frac{q(q-1)}{2}\int_0^T \|\theta(s)\|^2\,ds\right\}\right] \\ &\leqq \exp\left\{\frac{q(q-1)}{2}C\right\}\end{aligned} \tag{5.7}$$

from (2.6) with $1/p + 1/q = 1$. Now (5.5)-(5.7), in conjunction with the Hölder inequality, give

$$\begin{aligned}EU(X^{x,\pi}(T)) &= E_0[Z_0^{-1}(T)U(X^{x,\pi}(T))] \\ &\leqq (E_0 Z_0^{-q}(T))^{1/q}(E_0 U_+^p(X^{x,\pi}(T)))^{1/p} \\ &\leqq \exp\left\{\frac{(q-1)}{2}C\right\}(c_1 + c_2(e^L x)^{\delta p})^{1/p} < \infty\end{aligned}$$

for every $\pi \in \mathcal{A}(x)$, justifying Standing Assumption 5.1.

6. The complete market solution. The utility maximization problem of § 5 admits a simple solution in the case $m = d$ of a complete market; this solution was derived by Karatzas, Lehoczky, and Shreve [13] and independently by Cox and Huang [2], [3]. In this section we briefly review the pertinent results, both for easy reference and for later usage in the treatment of the incomplete market case.

For the purposes of this discussion, we need the following assumption.

Assumption 6.1. $E[\beta(T)Z_0(T)I(y\beta(T)Z_0(T))] < \infty$, for all $y \in (0, \infty)$.

Under it, the function $\mathcal{X}_0 : (0, \infty) \to (0, \infty)$ defined by

$$\mathcal{X}_0(y) \triangleq E[\beta(T)Z_0(T)I(y\beta(T)Z_0(T))], \qquad 0 < y < \infty \tag{6.1}$$

inherits from I the property of being a continuous, strictly decreasing mapping of $(0, \infty)$ onto itself, and so $\mathcal{X}_0$ has a (continuous, strictly decreasing) inverse $\mathcal{Y}_0$ from $(0, \infty)$ onto itself. We define

$$\xi_0^x \triangleq I(\mathcal{Y}_0(x)\beta(T)Z_0(T)). \tag{6.2}$$

Note that for every portfolio process $\pi \in \mathcal{A}(x)$, the supermartingale $\beta Z_0 X^{x,\pi}$ of (3.4) satisfies

$$E[\beta(T)Z_0(T)X^{x,\pi}(T)] \leqq x. \tag{6.3}$$

LEMMA 6.2. *The random variable* ξ_0^x *satisfies*

$$E[\beta(T)Z_0(T)\xi_0^x] = x, \tag{6.4}$$

$$E(U(\xi_0^x))^- < \infty, \tag{6.5}$$

and for every portfolio $\pi \in \mathscr{A}(x)$, *we have*

$$EU(X^{x,\pi}(T)) \leqq EU(\xi_0^x). \tag{6.6}$$

Proof. Equation (6.4) follows directly from the definitions of ξ_0^x and $\mathscr{Y}_0$. From (4.6) we have

$$\begin{aligned} U(\xi_0^x) &\geqq U(1) + \mathscr{Y}_0(x)\beta(T)Z_0(T)[\xi_0^x - 1] \\ &\geqq -|U(1)| - \mathscr{Y}_0(x)\beta(T)Z_0(T). \end{aligned} \tag{6.7}$$

But β is nonnegative and bounded almost surely and Z_0 is a nonnegative local martingale, thus a supermartingale. Therefore $E[\beta(T)Z_0(T)] < \infty$, and (6.5) follows. Now let π be a portfolio satisfying (5.1). From (4.6), (6.3), and (6.4) we have

$$\begin{aligned} EU(\xi_0^x) &\geqq E\{U(X^{x,\pi}(T)) + \mathscr{Y}_0(x)\beta(T)Z_0(T)[\xi_0^x - X^{x,\pi}(T)]\} \\ &\geqq EU(X^{x,\pi}(T)). \end{aligned}$$

□

From Lemma 6.2 it develops that if there exists a portfolio $\hat{\pi}$ such that $\xi_0^x = X^{x,\hat{\pi}}(T)$, then $\hat{\pi}$ is optimal. So far we have not used the assumption of market completeness; this assumption is used only in the *construction of the portfolio* $\hat{\pi}$ which finances ξ_0^x, a question that we now broach.

We begin with the positive martingale

$$M(t) \triangleq E[\beta(T)Z_0(T)\xi_0^x | \mathscr{F}_t]. \tag{6.8}$$

Being adapted to the Brownian filtration $\{\mathscr{F}_t\}$, M admits the stochastic integral representation

$$M(t) = x + \int_0^t \psi^*(s)\, dW(s) \tag{6.9}$$

for some $\{\mathscr{F}_t\}$-adapted process ψ satisfying $\int_0^T \|\psi(s)\|^2\, ds < \infty$ almost surely (e.g., Karatzas and Shreve [14, p. 184]). According to Itô's lemma

$$d\left(\frac{M(t)}{Z_0(t)}\right) = \frac{1}{Z_0(t)}(\psi(t) + M(t)\theta(t))^*\, dW_0(t),$$

and thus

$$\beta(T)\xi_0^x = \frac{M(T)}{Z_0(T)} = x + \int_0^T \frac{1}{Z_0(t)}(\psi(t) + M(t)\theta(t))^*\, dW_0(t).$$

We define

$$\hat{X}(t) \triangleq \frac{M(t)}{\beta(t)Z_0(t)} = \frac{1}{\beta(t)}\left[x + \int_0^t \frac{1}{Z_0(s)}(\psi(s) + M(s)\theta(s))^*\, dW_0(s)\right], \tag{6.10}$$

$$\hat{\pi}(t) \triangleq \frac{1}{\beta(t)Z_0(t)\hat{X}(t)}(\sigma^*(t))^{-1}(\psi(t) + M(t)\theta(t)), \tag{6.11}$$

and verify that $\hat{X}(0) = x$, $\hat{X}(T) = \xi_0^x$ as well as $d(\beta(t)\hat{X}(t)) = \beta(t)\hat{X}(t)\hat{\pi}^*(t)\sigma(t)\, dW_0(t)$ hold. A comparison with (3.3) shows that $\hat{X}(\cdot)$ is the wealth process corresponding to the portfolio $\hat{\pi}$: $\hat{X}(\cdot) \equiv X^{x,\hat{\pi}}(\cdot)$.

We have proved the following result.

THEOREM 6.3. *Let an initial wealth $x > 0$ be given. In a complete market* ($d = m$) *under Assumption* 6.1, *the portfolio $\hat{\pi}$ given by* (6.11) *is optimal. The resulting optimal terminal wealth is given by* (6.2).

Example 6.4 (*Logarithmic utility function*). Suppose $U(x) = \log x$. Then $\mathscr{X}_0(y) = 1/y$, $\mathscr{Y}_0(x) = 1/x$ and

$$\xi_0^x = x \exp\left\{\int_0^T \left(r(t) + \frac{1}{2}\|\theta(t)\|^2\right) dt + \int_0^T \theta^*(t)\, dW(t)\right\}. \tag{6.12}$$

Let $\hat{\pi}$ be given by

$$\hat{\pi}(t) \triangleq (\sigma(t)\sigma^*(t))^{-1}[b(t) - r(t)\mathbf{1}_m]. \tag{6.13}$$

From (3.4) we have $X^{x,\hat{\pi}}(T) = x/\beta(T)Z_0(T) = \xi_0^x$, so $\hat{\pi}$ is optimal and

$$V(x) = E[\log X^{x,\hat{\pi}}(T)] = \log x + E\int_0^T \left(r(t) + \frac{1}{2}\|\theta(t)\|^2\right) dt, \tag{6.14}$$

provided that this last expectation is finite (cf. Karatzas [12, §§ 9.3, 9.6]).

Example 6.5 (*Power utility function and deterministic model coefficients*). Suppose that $U(x) = x^\delta/\delta$, where $\delta < 1$, $\delta \neq 0$, and suppose that the processes r and θ are deterministic. Then $\exp\{(\delta/(1-\delta)) \int_0^t \theta^*(s)\, dW(s) - (\delta^2/2(1-\delta)^2) \int_0^t \|\theta(s)\|^2\, ds\}$ is a martingale with expectation equal to one (Karatzas and Shreve [14, Cor. 5.13, p. 199]), and from (6.1)

$$\begin{aligned}\mathscr{X}_0(y) &= y^{1/(\delta-1)} \exp\left\{\frac{\delta}{1-\delta}\int_0^T \left(r(s) + \frac{1}{2}\|\theta(s)\|^2\right) ds \cdot E \exp\left\{\frac{\delta}{1-\delta}\int_0^T \theta^*(s)\, dW(s)\right\}\right.\\ &= y^{1/(\delta-1)} \exp\left\{\frac{\delta}{1-\delta}\int_0^T m(s)\, ds\right\},\end{aligned}$$

where

$$m(t) \triangleq r(t) + \frac{1}{2(1-\delta)}\|\theta(t)\|^2. \tag{6.15}$$

It follows that $\mathscr{Y}_0(x) = x^{\delta-1} \exp\{\delta \int_0^T m(s)\, ds\}$, and

$$\xi_0^x = x \exp\left\{\int_0^T \left(r(t) + \frac{1-2\delta}{2(1-\delta)^2}\|\theta(t)\|^2\right) dt + \frac{1}{1-\delta}\int_0^T \theta^*(t)\, dW(t)\right\}. \tag{6.16}$$

Taking

$$\hat{\pi}(t) \triangleq \frac{1}{1-\delta}(\sigma(t)\sigma^*(t))^{-1}[b(t) - r(t)\mathbf{1}_m] \tag{6.17}$$

in (3.4), we obtain

$$\beta(t)Z_0(t)X^{x,\hat{\pi}}(t) = x \exp\left\{-\frac{\delta^2}{2(1-\delta)^2}\int_0^t \|\theta(s)\|^2\, ds + \frac{\delta}{1-\delta}\int_0^t \theta^*(s)\, dW(s)\right\},$$

from which follows $X^{x,\hat{\pi}}(T) = \xi_0^x$ and thereby the optimality of $\hat{\pi}$.

7. Fictitious completions of an incomplete market. The utility maximization problem of § 5 for an *incomplete market* ($d > m$) will be studied by the method of "fictitious completion." We will perform, in other words, the thought experiment of introducing $d - m$ additional stocks driven by the d-dimensional Brownian motion W, thus creating a fictitious complete market in which the utility maximization problem can be solved as in § 6. We will then try to determine appreciation rates for these additional stocks, so that the optimal portfolio in the resulting complete market *does not invest in the additional stocks at all,* i.e., is in $\mathscr{A}(x)$.

Following this program, we introduce an $\{\mathscr{F}_t\}$-progressively measurable, uniformly bounded, $(d-m)\times d$ matrix-valued process $\rho(t)$ whose rows, thought of as vectors in $\mathbb{R}^d$, are orthonormal and in the kernel of $\sigma(t)$, i.e., $\sigma(t)\rho^*(t)=0$. We also introduce an $\{\mathscr{F}_t\}$-progressively measurable, $(d-m)$-dimensional vector process $\mathscr{a}$ satisfying

$$\int_0^T \|\mathscr{a}(t)\|\,dt<\infty \quad \text{a.s.} \tag{7.1}$$

We create fictitious stocks with prices $S_i(t)$ governed by

$$dS_i(t)=S_i(t)\left[\mathscr{a}_i(t)\,dt+\sum_{j=1}^{d}\rho_{ij}(t)\,dW_j\,(t)\right], \qquad i=1,\cdots,d-m. \tag{7.2}$$

The matrix-valued process ρ will be held fixed throughout the remainder of the paper, but the process $\mathscr{a}$ will be considered as a parameter.

For the augmented stock appreciation rate vector $\tilde{b}\triangleq\left[\begin{smallmatrix} b \\ \mathscr{a}\end{smallmatrix}\right]$ and the augmented volatility matrix $\tilde{\sigma}\triangleq\left[\begin{smallmatrix} \sigma \\ \rho\end{smallmatrix}\right]$ we can define an augmented *relative risk* process

$$\tilde{\theta}(t)\triangleq\tilde{\sigma}^*(t)(\tilde{\sigma}(t)\tilde{\sigma}^*(t))^{-1}[\tilde{b}(t)-r(t)\mathbf{1}_m]=\theta(t)+\nu(t) \tag{7.3}$$

by analogy with (2.3), where

$$\nu(t)\triangleq\rho^*(t)[\mathscr{a}(t)-r(t)\mathbf{1}_{d-m}]. \tag{7.4}$$

Note that $\theta^*(t)\nu(t)=0$, and thus $\|\tilde{\theta}(t)\|^2=\|\theta(t)\|^2+\|\nu(t)\|^2$. It will be assumed that

$$\int_0^T \|\nu(t)\|^2\,dt<\infty \tag{7.5}$$

holds almost surely, so that (by analogy with (2.6) and (6.1)) we may define the exponential local martingale

$$\begin{aligned} Z_\nu(t)&\triangleq\exp\left\{-\int_0^t(\theta^*(s)+\nu^*(s))\,dW\,(s)-\frac{1}{2}\int_0^t(\|\theta(s)\|^2+\|\nu(s)\|^2)\,ds\right\}\\ &=1-\int_0^t Z_\nu(s)(\theta(s)+\nu(s))^*\,dW\,(s)\end{aligned} \tag{7.6}$$

and the function

$$\mathscr{X}_\nu(y)\triangleq E[\beta(T)Z_\nu(T)I(y\beta(T)Z_\nu(T))], \qquad 0<y<\infty. \tag{7.7}$$

If the condition

$$\mathscr{X}_\nu(y)<\infty \quad \forall y\in(0,\infty) \tag{7.8}$$

prevails, we may define $\mathscr{Y}_\nu$ to be the inverse of $\mathscr{X}_\nu$ and set

$$\xi_\nu^x\triangleq I(\mathscr{Y}_\nu(x)\beta(T)Z_\nu(T)) \tag{7.9}$$

by analogy with (6.2).

Remark 7.1. If the fictitious stocks introduced in this section were really available, then $EU(\xi_\nu^x)$ would be the maximal expected utility of final wealth (Theorem 6.3). Since these stocks are *not* available, we have

$$V(x)\triangleq\sup_{\pi\in\mathscr{A}(x)}EU(X^{x,\pi}(T))\leqq EU(\xi_\nu^x), \tag{7.10}$$

and equality holds if there exists a portfolio process $\pi\in\mathscr{A}(x)$ such that

$$X^{x,\pi}(T)=\xi_\nu^x \quad \text{a.s.}, \tag{7.11}$$

i.e., if the terminal wealth ξ_ν^x can be financed without investment in the fictitious stocks. In light of (7.10), such a π would be optimal for the problem of utility maximization in the incomplete market. In § 9 we shall discuss properties which π and ν must have in order to be related by (7.11).

8. A family of exponential local martingales. Let us denote by $S[0, T]$ the class of $\{\mathscr{F}_t\}$-adapted, $\mathbb{R}^d$-valued processes ψ satisfying

$$\int_0^T \|\psi(t)\|^2 \, dt < \infty \tag{8.1}$$

almost surely, and decompose $S[0, T]$ into the orthogonal subspaces

$$K(\sigma) \triangleq \{\nu \in S[0, T];\ \sigma(t)\nu(t) = 0, \quad \forall t \in [0, T], \text{a.s.}\}, \tag{8.2}$$

$$K^\perp(\sigma) \triangleq \{\varphi \in S[0, T];\ \varphi(t) \in \text{Range}\,(\sigma^*(t)), \forall t \in [0, T], \text{a.s.}\}. \tag{8.3}$$

Remark 8.1. The process θ of (2.3) belongs to $K^\perp(\sigma)$, whereas the process ν of (7.4) belongs to $K(\sigma)$. On the other hand, if $\nu \in K(\sigma)$ is given, then (7.4) can be solved for the appreciation rate vector α of the fictitious stocks, by taking this vector equal to

$$\alpha_\nu(t) \triangleq \rho(t)\nu(t) + r(t)\mathbf{1}_{d-m}. \tag{8.4}$$

Thus, the class $K(\sigma)$ provides a parameter space for fictitious completions of the incomplete market.

We will denote by $\mathfrak{M}_\nu$ the fictitious completion of the financial market by the additional stocks of (7.2), with $\rho(\cdot)$ fixed and $\alpha(\cdot) \equiv \alpha_\nu(\cdot)$, $\nu \in K(\sigma)$.

The associated family of exponential local martingales $\{Z_\nu\}_{\nu \in K(\sigma)}$, given by (7.6), will play a fundamental role in what follows.

LEMMA 8.2. *Consider the discounted stock price processes*

$$Q_i(t) \triangleq \beta(t)P_i(t), \qquad i = 1, \cdots, m.$$

Then for every $\nu \in K(\sigma)$, the processes $Z_\nu Q_i$ are local martingales under P.

Proof. It is seen from (2.2), (2.4) that

$$dQ_i(t) = Q_i(t)[(b_i(t) - r(t))\,dt + \sigma_i(t)\,dW(t)],$$

where $\sigma_i(t)$ is the ith row vector of the matrix $\sigma(t)$. It follows from this, (7.6), Itô's rule, and $\sigma(t)\nu(t) = 0$, that

$$d(Z_\nu(t)Q_i(t)) = Z_\nu(t)Q_i(t)[\sigma_i(t) - (\theta(t) + \nu(t))^*]\,dW(t). \qquad \square$$

PROPOSITION 8.3. *For any given $\pi \in \mathscr{A}(x)$, $\beta Z_\nu X^{x,\pi}$ is a local martingale under P for every $\nu \in K(\sigma)$; in particular,*

$$E[\beta(T)Z_\nu(T)X^{x,\pi}(T)] \leqq x \quad \forall \nu \in K(\sigma). \tag{8.5}$$

Proof. From (3.3), (2.5), and (7.6) follows the analogue

$$\beta(t)Z_\nu(t)X(t) = x + \int_0^t \beta(s)Z_\nu(s)X(s)[\sigma^*(s)\pi(s) - (\theta(s) + \nu(s))]^*\,dW(s) \tag{8.6}$$

of (3.4) for the process $X \equiv X^{x,\pi}$. This representation shows that $\beta Z_\nu X^{x,\pi}$ is a positive local martingale, hence a supermartingale, and (8.5) follows. $\square$

Remark 8.4. Suppose that π is a portfolio process, and that X is a continuous, $\{\mathscr{F}_t\}$-adapted process which satisfies (8.6) almost surely, *for some* $\nu \in K(\sigma)$. Then X is the wealth process corresponding to the initial endowment x and the portfolio process π, i.e., $X = X^{x,\pi}$. Indeed, apply Itô's rule to the product of the processes $\beta Z_\nu X$ and Λ_ν, where $\Lambda_\nu = Z_\nu^{-1}$ is easily seen from (7.6) to satisfy

$$d\Lambda_\nu(t) = \Lambda_\nu(t)[(\theta(t)+\nu(t))^* \, dW(t) + (\|\theta(t)\|^2 + \|\nu(t)\|^2)\, dt],$$

and obtain (3.3).

The following result provides a kind of "converse" to Proposition 8.3.

THEOREM 8.5. *Consider a positive, $\mathcal{F}_T$-measurable random variable B, for which there exists a process $\lambda \in K(\sigma)$ with*

$$E[\beta(T)Z_\nu(T)B] \leqq x = E[\beta(T)Z_\lambda(T)B] \quad \forall \nu \in K(\sigma). \tag{8.7}$$

Then there exists a portfolio $\pi \in \mathcal{A}(x)$, such that $X^{x,\pi}(T) = B$, almost surely.

Proof. Define a positive, $\{\mathcal{F}_t\}$-adapted process X via

$$\beta(t)Z_\lambda(t)X(t) = M(t) \triangleq E[\beta(T)Z_\lambda(T)B|\mathcal{F}_t], \qquad 0 \leqq t \leqq T. \tag{8.8}$$

Certainly $X(0) = x$, $X(T) = B$ almost surely, and the positive martingale M in (8.8) has $M(0) = x$. From the martingale representation theorem (Karatzas and Shreve [14, Problem 3.4.16, p. 184]), $M(t) = x + \int_0^t \varphi(s)\, dW(s)$ for some $\{\mathcal{F}_t\}$-adapted process φ satisfying $\int_0^T \|\varphi(t)\|^2\, dt < \infty$ almost surely. Since M is continuous and $M(t) > 0$ for all $t \in [0, T]$, we may define $\psi \in S[0, T]$ by $\psi(t) = -\varphi(t)/M(t)$. Then

$$\begin{aligned} M(t) &= x \exp\left\{ -\int_0^t \psi^*(s)\, dW(s) - \frac{1}{2}\int_0^t \|\psi(s)\|^2\, ds \right\} \\ &= x - \int_0^t M(s)\psi^*(s)\, dW(s), \qquad 0 \leqq t \leqq T. \end{aligned} \tag{8.9}$$

Decomposing ψ as $\psi = \psi_1 + \psi_2$ with $\psi_1 \in K^\perp(\sigma)$, $\psi_2 \in K(\sigma)$ and comparing (8.8), (8.9) with (8.6), it transpires that proving the theorem amounts to finding a portfolio π such that

$$-M(t)(\psi_1(t)+\psi_2(t)) = \beta(t)Z_\lambda(t)X(t)[\sigma^*(t)\pi(t) - (\theta(t)+\lambda(t))]. \tag{8.10}$$

This will certainly be possible, provided that

$$\psi_2(t) = \lambda(t) \quad dt \times dP \text{ a.e. on } [0, T] \times \Omega, \tag{8.11}$$

because we can take then π to satisfy $\sigma^*\pi = \theta - \psi_1 \in K^\perp(\sigma)$. Consequently, *we have to show that* (8.7) *implies* (8.11).

To this end, consider an arbitrary but fixed $\nu \in K(\sigma)$ and introduce the sequence of stopping times $\{\tau_n\}_{n=1}^\infty$ given by

$$\begin{aligned} \tau_n \triangleq T \wedge \inf\Big\{ t \in [0, T];\ M(t) \geqq n, \text{ or } &\int_0^t (\|\psi_1(s)\|^2 + \|\psi_2(s)\|^2 + \|\lambda(s)\|^2)\, ds \geqq n, \\ \text{or } \int_0^t \|\nu(s)\|^2\, ds \geqq n, &\text{ or } \left|\int_0^t \nu^*(s)\, dW(s)\right| \geqq n \Big\} \end{aligned} \tag{8.12}$$

for every $n \geqq 1$. Obviously, $\lim_{n\to\infty} \tau_n = T$ almost surely, and we denote $\nu_n(t) \triangleq \nu(t)1_{[0,\tau_n]}(t)$. Clearly, $\lambda + \varepsilon\nu_n \in K(\sigma)$ and

$$Z_{\lambda+\varepsilon\nu_n}(t) = Z_\lambda(t) \exp\left\{ -\varepsilon \int_0^{t\wedge\tau_n} \nu^*(s)(dW(s) + \lambda(s)\, ds) - \frac{\varepsilon^2}{2}\int_0^{t\wedge\tau_n} \|\nu(s)\|^2\, ds \right\}, \tag{8.13}$$

for every $\varepsilon \in (-1, 1)$, $n \geqq 1$. On the other hand, the definition of τ_n in (8.12) gives

$$e^{-3n|\varepsilon|} \leqq \frac{Z_{\lambda+\varepsilon\nu_n}(T)}{Z_\lambda(T)} \leqq e^{3n|\varepsilon|}, \qquad -1 < \varepsilon < 1. \tag{8.14}$$

It follows then quite easily (from (8.12)-(8.14) and the dominated convergence theorem) that (8.7) implies

$$0=\frac{\partial}{\partial\varepsilon}E\left[\beta(T)Z_{\lambda+\varepsilon\nu_n}(T)B\right]\bigg|_{\varepsilon=0}=E\left[\beta(T)\cdot\frac{\partial}{\partial\varepsilon}Z_{\lambda+\varepsilon\nu_n}(T)\bigg|_{\varepsilon=0}\cdot B\right] \tag{8.15}$$
$$=-E\left[\beta(T)Z_\lambda(T)B\int_0^{\tau_n}\nu^*(s)(dW(s)+\lambda(s)\,ds)\right],$$

or equivalently, in the notation of (8.8):

$$E\left[M(\tau_n)\int_0^{\tau_n}\nu^*(s)(dW(s)+\lambda(s)\,ds)\right]=0\quad\forall n\geqq 1. \tag{8.16}$$

Now Itô's rule, in conjunction with (8.9), gives

$$M(\tau_n)\int_0^{\tau_n}\nu^*(s)(dW(s)+\lambda(s)\,ds)$$
$$=\int_0^{\tau_n}M(t)\nu^*(t)(\lambda(t)-\psi_2(t))\,dt+\int_0^{\tau_n}M(t)\nu^*(t)\,dW(t) \tag{8.17}$$
$$-\int_0^{\tau_n}M(t)\left\{\int_0^t\nu^*(s)(dW(s)+\lambda(s)\,ds)\right\}(\psi_1(t)+\psi_2(t))^*\,dW(t).$$

From the definition of τ_n in (8.12) we see that the expectations of the two stochastic integrals in (8.17) are equal to zero. Substituting back into (8.16), we obtain

$$E\int_0^{\tau_n}M(t)\nu^*(t)(\lambda(t)-\psi_2(t))\,dt=0\quad\forall n\geqq 1. \tag{8.18}$$

The arbitrariness of $\nu\in K(\sigma)$ in (8.18) leads to (8.11). □

9. Equivalent optimality conditions in an incomplete market. The conclusions of § 7 were predicated on the assumption

$$\mathcal{X}_\nu(y)<\infty\quad\forall y\in(0,\infty), \tag{7.8}$$

but this condition often will not hold for all $\nu\in K(\sigma)$ (cf. § 13 (Appendix)). Accordingly, we restrict ourselves to the class

$$K_1(\sigma)\triangleq\{\nu\in K(\sigma);\ \nu\text{ satisfies (7.8)}\} \tag{9.1}$$

in what follows.

Remark 9.1. If Assumption 4.3 holds, and $\nu\in K(\sigma)$ satisfies $\mathcal{X}_\nu(y)<\infty$ for *some* $y\in(0,\infty)$, then $\nu\in K_1(\sigma)$. This can be verified easily, using (4.9).

For a fixed initial capital $x>0$, let $\hat\pi\in\mathcal{A}(x)$ be given, and consider the statement that $\hat\pi$ is optimal for the incomplete market maximization problem of § 5:

(A) OPTIMALITY OF $\hat\pi$. $EU(X^{x,\pi}(T))\leqq EU(X^{x,\hat\pi}(T))$, *for all* $\pi\in\mathcal{A}(x)$.

We will characterize condition (A) with the help of the following conditions (B)-(E). For a given $\lambda\in K_1(\sigma)$ recall the notation of (4.3), (7.9) and consider the following statements.

(B) FINANCIBILITY OF ξ^x_λ. *There exists a portfolio* $\hat\pi\in\mathcal{A}(x)$ *such that* $X^{x,\hat\pi}(T)\equiv\xi^x_\lambda$, *almost surely.*

(C) LEAST-FAVORABILITY OF λ. $EU(\xi^x_\lambda)\leqq EU(\xi^x_\nu)$, *for all* $\nu\in K_1(\sigma)$.

(D) DUAL OPTIMALITY OF λ. *For all* $\nu\in K_1(\sigma)$,

$$E\tilde U(\mathcal{Y}_\lambda(x)\beta(T)Z_\lambda(T))\leqq E\tilde U(\mathcal{Y}_\lambda(x)\beta(T)Z_\nu(T)).$$

(E) PARSIMONY OF λ. $E[\beta(T)Z_\nu(T)\xi^x_\lambda]\leqq x$, *for all* $\nu\in K_1(\sigma)$.

Our principal result of this section, Theorem 9.4, states that conditions (B)-(E) on λ are *equivalent* and they imply the existence of $\hat{\pi}$ satisfying (A), provided that Assumption 4.3 and $U(0)>-\infty$ hold. This latter restriction is rather severe, for it excludes the important special case of the logarithmic utility function $U(x)=\log x$. For this reason we also develop a somewhat more modest result, Theorem 9.3, which suffices for a complete treatment of the logarithmic case (Example 10.1).

But first, let us try to motivate the developments that follow by discussing the significance of conditions (B)-(E). While we do not present any proofs for the claimed equivalences in the discussion that follows, we offer some plausible arguments to the effect that conditions (A)-(E) are connected to one another.

Discussion 9.2. For any given $\lambda \in K_1(\sigma)$, ξ_λ^x is the optimal level of terminal wealth in the fictitiously completed market $\mathfrak{M}_\lambda$. When will it also be optimal in the original, incomplete market? Presumably, only when there exists a portfolio $\hat{\pi}$ *which invests in the original m stocks only* (i.e., $\hat{\pi} \in \mathcal{A}(x)$), such that $X^{x,\hat{\pi}} = \xi_\lambda^x$. In other words, condition (B) has then to hold, and condition (E) follows directly from (8.5). In particular, (E) says that the value of the contingent claim ξ_λ^x is at least as large in the fictitiously completed market $\mathfrak{M}_\lambda$ as in any other market $\mathfrak{M}_\nu$, $\nu \in K_1(\sigma)$. Note in this connection that, according to the definitions, $E[\beta(T)Z_\lambda(T)\xi_\lambda^x]=x$, for all $\lambda \in K_1(\sigma)$.

Furthermore, the terminal wealth ξ_λ^x can be financed by investing in the stocks of *any* other market $\mathfrak{M}_\nu$ (since, in fact, it can be financed by investing in the original m stocks). Thus we obtain the condition (C), which captures the "least favorable" character of λ.

Let us derive finally the condition (D), at least in the case $U(0)>-\infty$ (in which $\tilde{U}$ is bounded from below, thanks to Lemma 4.2, and thus the expectations in (D) are well defined). Indeed, by writing (4.7) with x replaced by ξ_λ^x and y replaced by $\mathcal{Y}_\lambda(x)\beta(T)Z_\nu(T)$, and taking expectations, we obtain

$$\begin{aligned} E\tilde{U}(\mathcal{Y}_\lambda(x)\beta(T)Z_\nu(T)) &\geqq E\tilde{U}(\mathcal{Y}_\lambda(x)\beta(T)Z_\lambda(T))+\mathcal{Y}_\lambda(x)\\ &\quad\cdot\{E[\beta(T)Z_\lambda(T)\xi_\lambda^x]-E[\beta(T)Z_\nu(T)\xi_\lambda^x]\}\\ &\geqq E\tilde{U}(\mathcal{Y}_\lambda(x)\beta(T)Z_\lambda(T)) \end{aligned}$$

from condition (E).

THEOREM 9.3. *Conditions* (B) *and* (E) *are equivalent, and imply* (C). *Furthermore, if* (B) *holds, then the portfolio* $\hat{\pi}$ *in* (B) *satisfies* (A).

Proof. (B)$\Rightarrow$(E) follows from Proposition 8.3.

(E)$\Rightarrow$(B) follows by letting $B\equiv\xi_\lambda^x$ in Theorem 8.5. Note that this theorem remains valid if in it $K(\sigma)$ is replaced by $K_1(\sigma)$. In order to see this, it suffices to observe that the processes $\lambda+\varepsilon\nu_n$ (appearing in (8.13)) belongs to $K_1(\sigma)$ for every $\varepsilon\in(-1,1)$, $n\geqq 1$, because from (8.14) and the fact that $\lambda \in K_1(\sigma)$:

$$\mathcal{X}_{\lambda+\varepsilon\nu_n}(y)\leqq e^{3n|\varepsilon|}\mathcal{X}_\lambda(y\,e^{-3n|\varepsilon|})<\infty \quad \forall y\in(0,\infty).$$

The last statement of the theorem follows from Remark 7.1.

(B)$\Rightarrow$(C) holds because the previous implication and (7.10) imply

$$EU(\xi_\lambda^x)=V(x)\leqq EU(\xi_\nu^x) \quad \forall \nu \in K_1(\sigma). \qquad \square$$

THEOREM 9.4. *Assume that* $U(0)>-\infty$ *holds. Then*

(i) *Conditions* (B)-(E) *are equivalent, and if* (B) *holds, then the portfolio* $\hat{\pi}$ *in* (B) *satisfies* (A);

(ii) *Conversely, if* $\hat{\pi}\in\mathcal{A}(x)$ *satisfies* (A), *then there exists a* $\lambda\in K_1(\sigma)$ *for which* (B)-(E) *hold, provided that Assumption* 4.3 *is also in force.*

Proof. In view of Theorem 9.3, we need discuss only the implications (C)⇒(D)⇒(B) and assertion (ii) under the appropriate conditions.

(C)⇒(D) For any given $y>0$ and $\nu\in K_1(\sigma)$, the convexity of $\tilde U$ yields

$$\frac{1}{|\varepsilon|}|\tilde U((y+\varepsilon)\beta(T)Z_\nu(T))-\tilde U(y\beta(T)Z_\nu(T))|\leqq\beta(T)Z_\nu(T)I\left(\frac{y}{2}\beta(T)Z_\nu(T)\right)\tag{9.2}$$

in conjunction with (4.4), for $\varepsilon>-y/2$, $\varepsilon\neq 0$. From the assumption $\nu\in K_1(\sigma)$, the random variable on the right-hand side of (9.2) has expectation equal to $\mathscr{X}_\nu(y/2)<\infty$, and the dominated convergence theorem shows that

$$\frac{d}{dy}E\tilde U(y\beta(T)Z_\nu(T))=-E[\beta(T)Z_\nu(T)I(y\beta(T)Z_\nu(T))]=-\mathscr{X}_\nu(y).\tag{9.3}$$

Therefore, for any given $x>0$, $\nu\in K_1(\sigma)$, the convex function

$$f_\nu(y)\triangleq E\tilde U(y\beta(T)Z_\nu(T))+xy,\qquad 0<y<\infty,\tag{9.4}$$

attains its minimum at $\mathscr{Y}_\nu(x)$, since $f'_\nu(y)=x-\mathscr{X}_\nu(y)$. But thanks to (C) and (4.3), we now have for any $y>0$ that

$$\begin{aligned}f_\nu(y)&\geqq f_\nu(\mathscr{Y}_\nu(x))=E[\tilde U(\mathscr{Y}_\nu(x)\beta(T)Z_\nu(T))\\&\quad+\mathscr{Y}_\nu(x)\beta(T)Z_\nu(T)\cdot I(\mathscr{Y}_\nu(x)\beta(T)Z_\nu(T))]\\&=EU(I(\mathscr{Y}_\nu(x)\beta(T)Z_\nu(T)))=EU(\xi^x_\nu)\geqq EU(\xi^x_\lambda)\\&=\cdots=f_\lambda(\mathscr{Y}_\lambda(x))\quad\forall\nu\in K_1(\sigma)\end{aligned}\tag{9.5}$$

and thus,

$$\begin{aligned}E\tilde U(\mathscr{Y}_\lambda(x)\beta(T)Z_\lambda(T))&=E[U(\xi^x_\lambda)-\mathscr{Y}_\lambda(x)\beta(T)Z_\lambda(T)\xi^x_\lambda]\\&=f_\lambda(\mathscr{Y}_\lambda(x))-x\mathscr{Y}_\lambda(x)\leqq f_\nu(\mathscr{Y}_\lambda(x))-x\mathscr{Y}_\lambda(x)\\&=E\tilde U(\mathscr{Y}_\lambda(x)\beta(T)Z_\nu(T)).\end{aligned}$$

(D)⇒(B) Repeat the proof of Theorem 8.5 up to (8.14), with $K(\sigma)$ replaced by $K_1(\sigma)$, (8.7) by (D), and B by ξ^x_λ. Everything then boils down to showing that the analogue

$$E\left[\beta(T)Z_\lambda(T)\xi^x_\lambda\int_0^{\tau_n}\nu^*(s)(dW(s)+\lambda(s)\,ds)\right]=0\tag{9.6}$$

of (8.15) can be obtained from the consequence of (D)

$$\frac{\partial}{\partial\varepsilon}E[\tilde U(\mathscr{Y}_\lambda(x)\beta(T)Z_{\lambda+\varepsilon\nu_n}(T))]|_{\varepsilon=0}=0,\tag{9.7}$$

since $\lambda+\varepsilon\nu_n\in K_1(\sigma)$ for every $\varepsilon\in(-1,1)$, $n\geqq 1$ (recall the argument in the proof of implication (E)⇒(B) in Theorem 9.3). Indeed, (9.6) follows formally from (9.7) by differentiating inside the expectation sign, and using (4.4), (7.9), (8.13).

For a rigorous justification, recall (8.12) and use the convexity of $\tilde U$ to obtain, for any given $y>0$:

$$\begin{aligned}&|\tilde U(y\beta(T)Z_{\lambda+\varepsilon\nu_n}(T))-\tilde U(y\beta(T)Z_\lambda(T))|\\&\quad\leqq y\beta(T)I(y\beta(T)\min\{Z_\lambda(T),Z_{\lambda+\varepsilon\nu_n}(T)\})\,|Z_{\lambda+\varepsilon\nu_n}(T)-Z_\lambda(T)|\\&\quad\leqq y\beta(T)(e^{3n|\varepsilon|}-1)Z_\lambda(T)I(y\beta(T)\,e^{-3n}Z_\lambda(T))\\&\quad\leqq K_n|\varepsilon|\cdot y\beta(T)Z_\lambda(T)I(y\beta(T)\,e^{-3n}Z_\lambda(T)),\end{aligned}\tag{9.8}$$

where $K_n \triangleq \sup_{0<\varepsilon<1}(e^{3n\varepsilon}-1)/\varepsilon$. The expectation of the right-hand side of (9.8) is equal to $yK_n|\varepsilon|$ times $E[\beta(T)Z_\lambda(T)I(ye^{-3n}\beta(T)Z_\lambda(T)] = \mathscr{X}_\lambda(ye^{-3n})$, a finite quantity by the assumption $\lambda \in K_l(\sigma)$.

On the other hand, the mean-value theorem implies that for each $\varepsilon \in (-1, 1)\backslash\{0\}$ there is a random variable γ_ε with values in $[0, 1]$, such that

$$\begin{aligned}
&\frac{1}{\varepsilon}[\tilde{U}(y\beta(T)Z_{\lambda+\varepsilon\nu_n}(T)) - \tilde{U}(y\beta(T)Z_\lambda(T))]\\
&\quad = y\beta(T)\frac{Z_{\lambda+\varepsilon\nu_n}(T)-Z_\lambda(T)}{\varepsilon}\cdot(\tilde{U}'(y\beta(T)\{Z_\lambda(T)+\gamma_\varepsilon(Z_{\lambda+\varepsilon\nu_n}(T)-Z_\lambda(T)\}))\\
&\quad = -I(y\beta(T)\{Z_\lambda(T)+\gamma_\varepsilon(Z_{\lambda+\varepsilon\nu_n}(T)-Z_\lambda(T))\})\cdot y\beta(T)Z_\lambda(T)\\
&\qquad \cdot\frac{1}{\varepsilon}\left[\exp\left\{-\varepsilon\int_0^{\tau_n}\nu^*(s)(dW(s)+\lambda(s)\,ds)-\frac{\varepsilon^2}{2}\int_0^{\tau_n}\|\nu(s)\|^2\,ds\right\}-1\right].
\end{aligned}$$

From this and (9.7), the conclusion (9.6) follows, thanks to the dominated convergence theorem, by letting $\varepsilon\downarrow 0$.

Proof of (ii). *Step* 1. Let $\hat{X}$ be the wealth process corresponding to the optimal portfolio $\hat{\pi}$. We have from (3.3)

$$\begin{aligned}
\beta(t)\hat{X}(t) &= x+\int_0^t \beta(s)\hat{X}(s)(\sigma^*(s)\hat{\pi}(s))^*\,dW_0(s)\\
&= x\exp\left\{\int_0^t \hat{\pi}^*(s)\sigma(s)\,dW_0(s)-\frac{1}{2}\int_0^t\|\sigma^*(s)\hat{\pi}(s)\|^2\,ds\right\}.
\end{aligned}\tag{9.9}$$

Now take a *bounded*, $\{\mathscr{F}_t\}$-progressively measurable portfolio process η with values in $\mathbb{R}^m$, and perform a small random perturbation of $\hat{\pi}$ according to

$$\pi_\varepsilon(t) \triangleq \hat{\pi}(t)+\varepsilon\eta(t)1_{\{t\leqq\tau_n\}},\tag{9.10}$$

where $-1<\varepsilon<1$, $\varepsilon\neq 0$, and

$$\begin{aligned}
\tau_n = T\wedge\inf\Big\{t\in[0, T];\ &\left|\int_0^t \eta(s)\sigma(s)\,dW_0(s)\right|\geqq n,\text{ or }\int_0^t\|\sigma^*(s)\hat{\pi}(s)\|^2\,ds\geqq n,\text{ or}\\
&\int_0^t\|\sigma^*(s)\eta(s)\|^2\,ds\geqq n,\text{ or }\int_0^t\|\theta(s)\|^2\,ds\geqq n,\text{ or }N(t)\geqq n,\text{ or}\\
&|A(t)|\geqq n,\text{ or }\int_0^t\|\psi(s)\|^2\,ds\geqq n\Big\}
\end{aligned}\tag{9.11}$$

(see (9.19), (9.20) below for the definitions of the processes N, A, and ψ). We also define the process $X_\varepsilon(\cdot)$ via

$$\begin{aligned}
\beta(t)X_\varepsilon(t) &\triangleq x\exp\left\{\int_0^t \pi_\varepsilon^*(s)\sigma(s)\,dW_0(s)-\frac{1}{2}\int_0^t\|\sigma^*(s)\pi_\varepsilon(s)\|^2\,ds\right\}\\
&= x+\int_0^t \beta(s)X_\varepsilon(s)\pi_\varepsilon^*(s)\sigma(s)\,dW_0(s),
\end{aligned}\tag{9.12}$$

and note that $X_\varepsilon(\cdot)\equiv X^{x,\pi_\varepsilon}(\cdot)$. Consequently, (A) gives

$$\frac{\partial}{\partial\varepsilon}EU(X_\varepsilon(T))|_{\varepsilon=0}=0.\tag{9.13}$$

A comparison of (9.9), (9.12) yields

$$X_\varepsilon(t)=\hat{X}(t)\exp\left\{\varepsilon\int_0^{t\wedge\tau_n}\eta^*(s)\sigma(s)\,d\hat{W}(s)-\frac{\varepsilon^2}{2}\int_0^{t\wedge\tau_n}\|\sigma^*(s)\eta(s)\|^2\,ds\right\},\tag{9.14}$$

where

$$\hat{W}(t)\triangleq W_0(t)-\int_0^t\sigma^*(s)\hat{\pi}(s)\,ds=W(t)+\int_0^t(\theta(s)-\sigma^*(s)\hat{\pi}(s))\,ds.\tag{9.15}$$

Then, at least formally, (9.13) and (9.14) lead to

$$E\left[U'(\hat{X}(T))\hat{X}(T)\int_0^{\tau_n}\eta^*(s)\sigma(s)\,d\hat{W}(s)\right]=0\quad\forall n\geqq 1.\tag{9.16}$$

Step 2. In order to justify (9.16) rigorously, observe from (9.14) and (9.11) that $e^{-3n|\varepsilon|}\leqq X_\varepsilon(T)/\hat{X}(T)\leqq e^{3n|\varepsilon|}$, almost surely, and from the concavity of U

$$\begin{aligned}\frac{1}{|\varepsilon|}|U(X_\varepsilon(T))-U(\hat{X}(T))|&\leqq U'(\min\{X_\varepsilon(T),\hat{X}(T)\})\left|\frac{X_\varepsilon(T)-\hat{X}(T)}{\varepsilon}\right|\\ &\leqq U'(e^{-3n|\varepsilon|}\hat{X}(T))\hat{X}(T)\frac{e^{3n|\varepsilon|}-1}{|\varepsilon|}\\ &\leqq[U(e^{-3n|\varepsilon|}\hat{X}(T))-U(0)]\,e^{3n}K_n\\ &\leqq e^{3n}K_n\cdot[U(\hat{X}(T))-U(0)],\end{aligned}\tag{9.17}$$

with K_n as in (9.8). The right-hand side of (9.17) has finite expectation, namely, $e^{3n}K_n(V(x)-U(0))$. On the other hand, the mean value theorem implies the existence of a random variable γ_ε with values in $[0, 1]$, such that

$$\begin{aligned}\frac{1}{\varepsilon}[U(X_\varepsilon(T))-U(\hat{X}(T))]&=\frac{1}{\varepsilon}(X_\varepsilon(T)-\hat{X}(T))\cdot U'(\hat{X}(T)+\gamma_\varepsilon\{X_\varepsilon(T)-\hat{X}(T)\})\\ &=U'(\hat{X}(T)+\gamma_\varepsilon\{X_\varepsilon(T)-\hat{X}(T)\})\hat{X}(T)\\ &\quad\cdot\frac{1}{\varepsilon}\left[\exp\left\{\varepsilon\int_0^{\tau_n}\eta^*(s)\sigma(s)\,d\hat{W}(s)\right.\right.\\ &\quad\left.\left.-\frac{\varepsilon^2}{2}\int_0^{\tau_n}\|\sigma^*(s)\eta(s)\|^2\,ds\right\}-1\right].\end{aligned}$$

It is clear now that (9.16) follows from this expansion, (9.13), and the dominated convergence theorem, by letting $\varepsilon\downarrow 0$.

Step 3. Now proving (B) amounts to finding $\lambda\in K_1(\sigma)$ such that $\hat{X}(T)=I(\mathcal{Y}_\lambda(x)\beta(T)Z_\lambda(T))$, or equivalently

$$U'(\hat{X}(T))=\mathcal{Y}_\lambda(x)\beta(T)Z_\lambda(T).\tag{9.18}$$

We will show that (9.16) leads to a "natural" candidate process $\lambda\in K(\sigma)$, which is actually in $K_1(\sigma)$ and for which (9.18) is then shown to hold (Step 4).

Consider the process

$$\begin{aligned}A(t)&\triangleq\int_0^t\eta^*(s)\sigma(s)\,d\hat{W}(s)\\ &=\int_0^t\eta^*(s)\sigma(s)\,dW(s)+\int_0^t\eta^*(s)\sigma(s)[\theta(s)-\sigma(s)\hat{\pi}(s)]\,ds,\end{aligned}\tag{9.19}$$

as well as the positive martingale

$$N(t) \triangleq E[U'(\hat{X}(T))\hat{X}(T)|\mathcal{F}_t] = y_0 + \int_0^t N(s)\psi^*(s)\,dW(s), \tag{9.20}$$

where $y_0 = EN(T)$ and ψ is some process in $S[0, T]$, constructed by the argument preceding (8.9). Using Standing Assumption 5.1 and $U(0) > -\infty$, note that $y_0 = EN(T) \leqq E[U(\hat{X}(T)) - U(0)] < \infty$. Obviously (9.16) amounts to $E[N(\tau_n)A(\tau_n)] = 0$; on the other hand, we have from (9.19) and (9.20) that

$$\begin{aligned} N(\tau_n)A(\tau_n) = \int_0^{\tau_n} & N(t)\eta^*(t)\sigma(t)[\psi(t) + \theta(t) - \sigma^*(t)\hat{\pi}(t)]\,dt \\ & + \int_0^{\tau_n} N(t)[\sigma^*(t)\eta(t) + A(t)\psi(t)]^*\,dW(t). \end{aligned}$$

From the definition of τ_n in (9.11), the stochastic integral has zero expectation, and thus $E[N(\tau_n)A(\tau_n)] = 0$ leads to

$$E\int_0^{\tau_n} N(t)\eta^*(t)\sigma(t)[\psi(t) + \theta(t) - \sigma^*(t)\hat{\pi}(t)]\,dt = 0 \quad \forall n \geqq 1 \tag{9.21}$$

for arbitrary η as described above. Because $\tau_n \nearrow T$ almost surely as $n \to \infty$, we obtain that

$$\lambda \triangleq \sigma^*\hat{\pi} - (\psi + \theta) \tag{9.22}$$

belongs to $K(\sigma)$. For this choice of λ, the exponential local martingale Z_λ of (7.6) becomes

$$\begin{aligned} Z_\lambda(t) &= \exp\left\{-\int_0^t (\theta(s) + \lambda(s))^*\,dW(s) - \frac{1}{2}\int_0^t \|\theta(s) + \lambda(s)\|^2\,ds\right\} \\ &= \exp\left\{\int_0^t (\psi(s) - \sigma^*(s)\hat{\pi}(s))^*\,dW(s) - \frac{1}{2}\int_0^t \|\psi(s) - \sigma^*(s)\hat{\pi}(s)\|^2\,ds\right\}. \end{aligned} \tag{9.23}$$

Step 4. Finally, we justify $\lambda \in K_1(\sigma)$ and (9.18). From (9.20) and (9.9) and $\sigma\lambda = 0$, it follows that

$$\begin{aligned} U'(\hat{X}(T)) &= \beta(T)\frac{N(T)}{\beta(T)\hat{X}(T)} \\ &= \beta(T)\frac{y_0}{x}\frac{\exp\{\int_0^T \psi^*(s)\,dW(s) - \frac{1}{2}\int_0^T \|\psi(s)\|^2\,ds\}}{\exp\{\int_0^T (\sigma^*(s)\hat{\pi}(s))^*\,dW_0(s) - \frac{1}{2}\int_0^T \|\sigma^*(s)\hat{\pi}(s)\|^2\,ds\}} \\ &= \frac{y_0}{x}\beta(T)Z_\lambda(T), \end{aligned} \tag{9.24}$$

thanks to (9.23) and (9.22). It remains to show that $\lambda \in K_1(\sigma)$ and $y_0 = x\mathcal{Y}_\lambda(x)$. In order to see this, apply $I(\cdot)$ to both sides of (9.24), take expectations, and use (9.24) again to obtain

$$\begin{aligned} \mathcal{X}_\lambda\left(\frac{y_0}{x}\right) &= E\left[\beta(T)Z_\lambda(T)I\left(\frac{y_0}{x}\beta(T)Z_\lambda(T)\right)\right] = E[\beta(T)Z_\lambda(T)\hat{X}(T)] \\ &= \frac{x}{y_0}E[U'(\hat{X}(T))\hat{X}(T)] = \frac{x}{y_0}EN(T) = x < \infty. \end{aligned}$$

From Remark 9.1 we have $\lambda \in K_1(\sigma)$, and $y_0 = x\mathcal{Y}_\lambda(x)$ follows. □

10. Examples in an incomplete market. In the examples of this section, we assume $m<d$ and produce the optimal portfolio $\hat{\pi}$ and the process $\lambda \in K_1(\sigma)$ satisfying conditions (B), (C), and (E). In Example 10.2, λ also satisfies (D).

Example 10.1 (*Logarithmic utility function*). Suppose $U(x)=\log x$. Then $\mathscr{X}_\nu(y)=1/y$, $\mathscr{Y}_\nu(x)=1/x$, for all $\nu \in K(\sigma)$, and $\xi_\nu^x = x/\beta(T)Z_\nu(T)$. The process $\lambda \equiv 0$ satisfies (E), because

$$E[\beta(T)Z_\nu(T)\xi_0^x]=xE\left[\exp\left\{-\int_0^T \nu^*(s)\,dw\,(s)-\frac{1}{2}\int_0^T \|\nu(s)\|^2\,ds\right\}\right] \leqq x \quad \forall \nu \in K(\sigma).$$

The last inequality follows from the fact that $\exp\{-\int_0^t \nu^*(s)\,dW\,(s)-\frac{1}{2}\int_0^t \|\nu(s)\|^2\,ds\}$, being a nonnegative local martingale, must be a supermartingale. According to Theorem 9.3, the optimal portfolio process $\hat{\pi}$ must satisfy $X^{x,\hat{\pi}}(T)=\xi_0^x$, and this $\hat{\pi}$ was determined in (6.13) of Example 6.4. The value function $V(x)$ is given by the expression (6.14), and it is finite for every $x \in (0,\infty)$ if $E\int_0^T \|\theta(t)\|^2\,dt<\infty$ (recall Standing Assumption 5.1). From (7.4) we see that $\lambda \equiv 0$ corresponds to completion of the market by stocks whose appreciation rates are equal to the interest rate. *With a logarithmic utility function, the agent will not use such stocks even for hedging purposes.*

Example 10.2 (*Power utility function and totally unhedgeable market coefficients*). Suppose $U(x)=x^\delta/\delta$, where $\delta<1$, $\delta \neq 0$. Suppose that the volatility matrix $\sigma(t)$ has the form $\sigma(t)=[\check{\sigma}(t), 0]$, where $\check{\sigma}(t)$ is an $m \times m$, nonsingular matrix for all $t \in [0, T]$, almost surely. Decompose W into $\check{W}(t)=(W_1(t),\cdots, W_m(t))^*$ and $\mathring{W}(t)=(W_{m+1}(t),\cdots, W_d(t))^*$, and let $\{\check{\mathscr{F}}_t\}$ and $\{\mathring{\mathscr{F}}_t\}$ be the augmentations under P of the (independent) filtrations generated by $\check{W}$ and $\mathring{W}$, respectively. Assume that the processes r, b, and $\check{\sigma}$ are adapted to $\{\mathring{\mathscr{F}}_t\}$, a situation we refer to as *totally unhedgeable market coefficients* because the stock prices are driven solely by $\check{W}$:

$$dP_i(t)=P_i(t)\left[b_i(t)\,dt+\sum_{j=1}^m \check{\sigma}_{ij}(t)d\check{W}_j(t)\right], \qquad i=1,\cdots, m.$$

We show that under these conditions, the portfolio process given by (6.17) is optimal. In the present context, this process is random and $\{\mathring{\mathscr{F}}_t\}$-adapted, rather than deterministic as in Example 6.5.

To verify the above assertion, we note first that $\theta^*(t)=[\check{\theta}^*(t), 0]$, where

$$\check{\theta}(t) \triangleq \check{\sigma}(t)(\sigma(t)\check{\sigma}^*(t))^{-1}[b(t)-r(t)\mathbf{1}].$$

We note also that the processes $\lambda \in K(\sigma)$ are of the form $\lambda^*(t)=[0, \mathring{\lambda}^*(t)]$, where $\mathring{\lambda}(t)$ is $(d-m)$-dimensional. With $m(\cdot)$ given by (6.15), we define

$$A \triangleq E\exp\left\{\delta\int_0^T m(t)\,dt\right\}.$$

The differential of the positive $\{\mathring{\mathscr{F}}_t\}$-martingale $N(t) \triangleq E[\exp\{\delta\int_0^T m(s)\,ds\}|\mathring{\mathscr{F}}_t\}$ has a representation as $dN(t)=-N(t)\mathring{\lambda}^*(t)\,d\mathring{W}\,(t)$, where $\lambda=\begin{bmatrix} 0 \\ \mathring{\lambda} \end{bmatrix} \in K(\sigma)$ (see the argument leading to (8.9) for a justification). Therefore

$$\exp\left\{\delta\int_0^T m(t)\,dt\right\}=N(T)=A\exp\left\{-\int_0^T \mathring{\lambda}^*(t)\,d\mathring{W}\,(t)-\frac{1}{2}\int_0^T \|\mathring{\lambda}(t)\|^2\,dt\right\}. \tag{10.1}$$

We may assume without loss of generality that $\check{W}$ is the coordinate mapping process $\check{W}(t, \check{\omega})=\check{\omega}(t)$ defined on $\check{\Omega} \triangleq C([0, T], \mathbb{R}^m)$, the space of continuous functions from $[0, T]$ to $\mathbb{R}^m$, and $\mathring{W}$ is the coordinate mapping process on $\mathring{\Omega} \triangleq C([0, T], \mathbb{R}^{d-m})$. Then $\Omega = \check{\Omega} \times \mathring{\Omega}$, and P is the product of m-dimensional Wiener measure $\check{P}$ on $\check{\Omega}$ and $(d-m)$-dimensional Wiener measure $\mathring{P}$ on $\mathring{\Omega}$. Abusing notation slightly, we regard the $\{\mathring{\mathscr{F}}_t\}$-adapted process $\check{\theta}$ as a process on $\mathring{\Omega}$. For $\mathring{P}$-almost every $\mathring{\omega} \in \mathring{\Omega}$, we have $\int_0^T \|\check{\theta}^*(s, \mathring{\omega})\|^2\, ds < \infty$, and thus for fixed $\mathring{\omega}$, the process

$$(t, \check{w}) \mapsto \exp\left\{\frac{\delta}{1-\delta} \int_0^t \check{\theta}(s, \mathring{\omega})\, d\check{W}(s, \check{\omega}) - \frac{\delta^2}{2(1-\delta)^2} \int_0^t \|\check{\theta}(s, \mathring{\omega})\|^2\, ds\right\}$$

is an $\{\check{\mathscr{F}}_t\}$-martingale on $\check{\Omega}$ under $\check{P}$, with expectation equal to one (Karatzas and Shreve [14, Cor. 5.13, p. 199]). Consequently,

$$\begin{aligned} &E\left[\exp\left\{\frac{\delta}{1-\delta} \int_0^T \check{\theta}^*(s)\, d\check{W}(s)\right\} \middle| \mathring{\mathscr{F}}_T\right](\check{\omega}, \mathring{\omega}) \\ (10.2) \qquad &= \int_{\check{\Omega}} \exp\left\{\frac{\delta}{1-\delta} \int_0^T \check{\theta}^*(s, \mathring{\omega})\, d\check{W}(s, \check{\omega})\right\} P(d\check{\omega}) \\ &= \exp\left\{\frac{\delta^2}{2(1-\delta)^2} \int_0^T \|\check{\theta}(s, \mathring{\omega})\|^2\, ds\right\}. \end{aligned}$$

From (10.2) and (10.1) we have

$$\begin{aligned} \mathscr{X}_\lambda(y) = y^{1/(\delta-1)} E\Bigg[\exp\Bigg\{&\frac{\delta}{1-\delta} \int_0^T \left(r(s) + \frac{1}{2}\|\check{\theta}(s)\|^2 + \frac{1}{2}\|\mathring{\lambda}(s)\|^2\right) ds \\ &+ \frac{\delta}{1-\delta} \int_0^T \mathring{\lambda}^*(s)\, d\mathring{W}(s)\Bigg\} E\left[\exp\left\{\frac{\delta}{1-\delta} \int_0^T \check{\theta}^*(s)\, d\check{W}(s)\right\} \middle| \mathring{\mathscr{F}}_T\right]\Bigg] \\ = y^{1/(\delta-1)} E\Bigg[&\exp\left\{\frac{\delta}{1-\delta} \int_0^T \left(m(s)\, ds + \frac{1}{2}\|\mathring{\lambda}(s)\|^2\right) ds + \frac{\delta}{1-\delta} \int_0^T \mathring{\lambda}^*(s)\, d\mathring{W}(s)\right\}\Bigg] \\ = A^{\delta/(1-\delta)} y^{1/(\delta-1)} &E\left[\exp\left\{\delta \int_0^T m(s)\, ds\right\}\right] = \left(\frac{y}{A}\right)^{1/(\delta-1)} \end{aligned}$$

It follows that $\lambda \in K_1(\sigma)$, $\mathscr{Y}_\lambda(x) = Ax^{\delta-1}$, and using (10.1) we obtain

$$\begin{aligned} \xi_\lambda^x = (\mathscr{Y}_\lambda(x)\beta(T)Z_\lambda(T))^{1/(\delta-1)} = x \exp\Bigg\{&\int_0^T \left(r(t) + \frac{1-2\delta}{2(\delta-1)^2}\|\theta(t)\|^2\right) dt \\ &+ \frac{1}{1-\delta} \int_0^T \theta^*(t)\, dW(t)\Bigg\}. \end{aligned}$$

Just as in Example 6.5, we conclude from (3.4) that $X^{x,\hat{\pi}}(T) = \xi_\lambda^x$, where $\hat{\pi}$ is given by (6.17).

Remark 10.3. An important unresolved question is whether there are simple, widely applicable conditions that guarantee that for the process λ satisfying conditions (B)-(E) of § 9, the nonnegative local martingale Z_λ is actually a martingale. (See, however, the note following the references.) In Example 10.1 we have

$$Z_\lambda(t) = Z_0(t) = \exp\left\{-\int_0^t \theta^*(s)\, dW(s) - \frac{1}{2}\int_0^t \|\theta(s)\|^2\, ds\right\},$$

so we must assume at least that Z_0 is a martingale in order to conclude that Z_λ is. In Example 10.2 a computation similar to (10.2) reveals that

$$\begin{aligned} EZ_\lambda(T) &= E\bigg[\exp\bigg\{-\int_0^T \mathring{\lambda}^*(t)\,d\mathring{W}(t) - \frac{1}{2}\int_0^T (\|\check{\theta}(t)\|^2 + \|\mathring{\lambda}(t)\|^2)\,dt\bigg\} \\ &\qquad \cdot E\bigg[\exp\bigg\{-\int_0^T \check{\theta}^*(t)\,d\check{W}(t)\bigg\}\bigg|\mathring{\mathcal{F}}_T\bigg]\bigg] \\ &= E\bigg[\exp\bigg\{-\int_0^T \mathring{\lambda}^*(t)\,d\mathring{W}(t) - \frac{1}{2}\int_0^T \|\mathring{\lambda}(t)\|^2\,dt\bigg\}\bigg]. \end{aligned}$$

Taking expectations in (10.1) and recalling the definition of A, we see that $EZ_\lambda(T) = 1$. This is enough to ensure that Z_λ is a martingale (Karatzas and Shreve [14, p. 198]). We have so far been unable to produce an example in which Z_0 is a martingale but Z_λ is not.

11. Duality. We henceforth impose the following assumption.

Assumption 11.1. $U(0) > -\infty$.

In addition to the original, or "primal," optimization problem

$$V(x) = \sup_{\pi \in \mathcal{A}(x)} J(x; \pi), \qquad J(x; \pi) \triangleq EU(X^{x,\pi}(T)) \tag{11.1}$$

of § 5, we shall consider in what follows the *dual optimization problem* for $y \in (0, \infty)$, namely,

$$\tilde{V}(y) = \inf_{\nu \in K(\sigma)} \tilde{J}(y; \nu), \qquad \tilde{J}(y; \nu) \triangleq E\tilde{U}(y\beta(T)Z_\nu(T)). \tag{11.2}$$

This problem will have a value function $\tilde{V}: (0, \infty) \to \mathbb{R}$ under the following assumption, which will also remain in force for the remainder of the paper.

Assumption 11.2. For every $y \in (0, \infty)$, there exists $\nu \in K(\sigma)$ such that

$$\tilde{J}(y; \nu) < \infty. \tag{11.3}$$

See Remark 11.9 in connection with this assumption.

For arbitrary $x > 0$, $y > 0$, $\pi \in \mathcal{A}(x)$, and $\nu \in K(\sigma)$ it follows from (4.3) that

$$U(X^{x,\pi}(T)) \leqq \tilde{U}(y\beta(T)Z_\nu(T)) + y\beta(T)Z_\nu(T)X^{x,\pi}(T), \tag{11.4}$$

with equality if and only if

$$X^{x,\pi}(T) = I(y\beta(T)Z_\nu(T)). \tag{11.5}$$

By taking expectations in (11.4) and recalling Proposition 8.3, we obtain

$$J(x; \pi) \leqq \tilde{J}(y; \nu) + xy, \tag{11.6}$$

with equality prevailing if and only if (11.5) and $x = \mathcal{X}_\nu(y)$ hold. In particular, it follows from (11.6) that

$$V(x) \leqq \tilde{V}(y) + xy \quad \forall x > 0,\ y > 0. \tag{11.7}$$

Remark 11.3. Suppose that for some given $x > 0$ and $y > 0$, there exist $\hat{\pi}_x \in \mathcal{A}(x)$ and $\lambda_y \in K(\sigma)$ such that

$$J(x; \hat{\pi}_x) = \tilde{J}(y; \lambda_y) + xy. \tag{11.8}$$

Then $\hat{\pi}_x$ achieves the supremum in (11.1), and λ_y achieves the infimum in (11.2).

PROPOSITION 11.4. *Under Assumptions* 11.1 *and* 11.2, *suppose that, for a given* $y>0$, *there is an optimal process* $\lambda_y \in K_1(\sigma)$ *for the dual problem of* (11.2). *Then there exists an optimal portfolio* $\hat{\pi}_x \in \mathcal{A}(x)$ *for the primal problem of* (11.1) *with* $x=\mathcal{X}_{\lambda_y}(y)$, *and we have*

$$\tilde{V}(y)=\sup_{\xi>0}[V(\xi)-y\xi]. \tag{11.9}$$

In particular, $\tilde{V}(\cdot)$ *is convex.*

Proof. The optimality of λ_y gives

$$E[\tilde{U}(\mathcal{Y}_{\lambda_y}(x)\beta(T)Z_{\lambda_y}(T))]\leqq E[\tilde{U}(\mathcal{Y}_{\lambda_y}(x)\beta(T)Z_\nu(T))] \quad \forall \nu\in K(\sigma) \tag{11.10}$$

for this particular $x=\mathcal{X}_{\lambda_y}(y)$. This is Condition (D) of § 9. Theorem 9.4(i) shows the existence of a portfolio $\hat{\pi}_x \in \mathcal{A}(x)$, which is optimal for the primal problem and satisfies

$$X^{x,\hat{\pi}_x}(T)=I(y\beta(T)Z_{\lambda_y}(T)).$$

We conclude that (11.8) prevails (i.e., (11.6) holds as an equality with $\pi=\hat{\pi}_x$, $\nu=\lambda_y$), and thus

$$\tilde{V}(y)=\tilde{J}(y;\lambda_y)=J(x;\hat{\pi}_x)-xy=V(x)-xy\leqq\sup_{\xi>0}[V(\xi)-y\xi].$$

The inequality in the opposite direction follows directly from (11.7), and the duality relationship (11.9) is established. □

Assumption 11.5. Suppose that the dual problem of (11.2) admits an optimal process $\lambda_y\in K_1(\sigma)$, *for every* $y>0$.

A sufficient condition (Theorem 12.3) for the fulfillment of Assumption 11.5 will be given in the next section. Under the Assumption 11.5, (11.9) holds for all $y>0$, and the following question arises. *Under what conditions can we guarantee that, for every given* $x>0$, *there exists an optimal portfolio* $\hat{\pi}_x$ *for* (11.1)? According to Proposition 11.4, this will happen if for every $x>0$ we can find a real number $y(x)>0$ such that

$$x=\mathcal{X}_{\lambda_{y(x)}}(y(x)). \tag{11.11}$$

PROPOSITION 11.6. *Suppose that the conditions of Proposition* 11.4 *hold, as well as Assumptions* 11.5 *and* 4.3 *and* $U(\infty)=\infty$. *Then for every* $x>0$, *there exists a real number* $y(x)>0$ *that achieves* $\inf_{y>0}[\tilde{V}(y)+xy]$; *this number satisfies* (11.11) *as well.*

Proof. From (11.3), Jensen's inequality, the supermartingale property of Z_ν, and the decrease of $\tilde{U}$, we have

$$\tilde{J}(y;\nu)\geqq\tilde{U}(yE[\beta(T)Z_\nu(T)])\geqq\tilde{U}(ye^L EZ_\nu(T))\geqq\tilde{U}(ye^L) \quad \forall\nu\in K(\sigma) \tag{11.12}$$

for the constant $L>0$ of Standing Assumption 2.1. Therefore, $\tilde{V}(y)\geqq\tilde{U}(ye^L)$ holds for every $y\in(0,\infty)$, and $\tilde{V}(0)\triangleq\lim_{y\downarrow 0}\tilde{V}(y)\geqq\tilde{U}(0)=U(\infty)=\infty$ (Lemma 4.2).

Consequently, for any given $x>0$, the convex function $f_x(y)=\tilde{V}(y)+xy$, $0<y<\infty$ satisfies $f_x(0+)=f_x(\infty)=\infty$, and thus attains its infimum on $(0,\infty)$ at some point $y(x)>0$. Now by the Assumption 11.5, there exists a process $\lambda_{y(x)}\in K_1(\sigma)$ such that $\tilde{V}(y(x))=\tilde{J}(y(x);\lambda_{y(x)})$, and we have

$$\begin{aligned}\inf_{\eta>0}[\eta y(x)x+\tilde{J}(\eta y(x);\lambda_{y(x)})]&=\inf_{y>0}[xy+\tilde{J}(y;\lambda_{y(x)})]\\&\geqq\inf_{y>0}[xy+\tilde{V}(y)]=xy(x)+\tilde{V}(y(x)).\end{aligned}$$

In other words, with the notation

$$G_y(u)\triangleq\tilde{J}(uy;\lambda_y)=E\tilde{U}(uy\beta(T)Z_{\lambda_y}(T)), \qquad 0<u<\infty, \tag{11.13}$$

the function

$$D_x(u) \triangleq uxy(x) + G_{y(x)}(u), \qquad 0 < u < \infty \tag{11.14}$$

achieves its infimum at $u = 1$.

From these considerations and Lemma 11.7 below, it transpires that

$$D'_x(1) = xy(x) + G'_{y(x)}(1) = xy(x) - y(x)\mathscr{X}_{\lambda_{y(x)}}(y(x))$$

is equal to zero, and thus (11.11) holds. □

LEMMA 11.7. *Under the conditions of Proposition* 11.6, *the function* $G_y(\cdot)$ *of* (11.13) *is well defined and finite on* $(0, \infty)$ *for any given* $0 < y < \infty$, *and satisfies*

$$G'_y(1) = -y\mathscr{X}_{\lambda_y}(y). \tag{11.15}$$

Proof. Since $\tilde{U}(\infty) = U(0) > -\infty$ by assumption and by Lemma 4.2, we have from (4.4) that

$$\tilde{U}(y) - \tilde{U}(\infty) = -\int_y^\infty \tilde{U}'(\xi)\, d\xi = \int_y^\infty I(\xi)\, d\xi, \qquad 0 < y < \infty.$$

Thus, for any given $\alpha \in (0, 1)$, it follows with the help of (4.9) that

$$\tilde{U}(\alpha y) - \tilde{U}(\infty) = \int_{\alpha y}^\infty I(\xi)\, d\xi = \alpha \int_y^\infty I(\alpha\eta)\, d\eta$$

$$\leqq \alpha\gamma \int_y^\infty I(\eta)\, d\eta = \alpha\gamma[\tilde{U}(y) - \tilde{U}(\infty)], \qquad 0 < y < \infty$$

for a suitable constant $\gamma \in (1, \infty)$. Consequently, for any given $y \in (0, \infty)$,

$$E\tilde{U}(\alpha y\beta(T)Z_{\lambda_y}(T)) \leqq \alpha\gamma E\tilde{U}(y\beta(T)Z_{\lambda_y}(T)) + (1 - \alpha\gamma)\tilde{U}(\infty)$$

$$= \alpha\gamma\tilde{J}(y; \lambda_y) + (1 - \alpha\gamma)U(0) < \infty.$$

Since $\alpha \in (0, 1)$ is arbitrary,

$$E\tilde{U}(uy\beta(T)Z_{\lambda_y}(T)) < \infty \tag{11.16}$$

holds for every $u \in (0, 1]$. But the function $\tilde{U}(\cdot)$ is decreasing, so (11.16) also holds for every $u > 1$.

Now use the convexity of $\tilde{U}$, the dominated convergence theorem, (4.4), and the fact that $\lambda_y \in K_1(\sigma)$, to justify the computations

$$G'_y(1) = yE[\beta(T)Z_{\lambda_y}(T)\tilde{U}'(y\beta(T)Z_{\lambda_y}(T))]$$

$$= -yE[\beta(T)Z_{\lambda_y}(T)I(y\beta(T)Z_{\lambda_y}(T))] = -y\mathscr{X}_{\lambda_y}(y),$$

which leads to (11.15). □

It just remains now to put Propositions 11.4 and 11.6 together, in order to obtain the following existence result for the primal problem (11.1).

THEOREM 11.8. *Suppose that Assumptions* 4.3, 11.1, 11.2, *and* 11.5 *hold and* $U(\infty) = \infty$. *Then for any given level* $x > 0$ *of initial capital, there exists an optimal portfolio* $\hat{\pi}_x \in \mathscr{A}(x)$ *for the utility maximization problem of* § 5.

In other words, under appropriate conditions, in order to obtain the existence of an optimal portfolio it is sufficient to deal with the existence of a solution $\lambda_y \in K_1(\sigma)$ for the dual problem (11.2). We will do that in the next section.

Remark 11.9. Assumption 11.2 is satisfied if (5.3) and (5.4) hold. Indeed, it is not hard to check that condition (5.4) and definition (4.3) lead to

$$\tilde{U}(y) \leqq k_1 + k_3 y^{-\alpha} \quad \forall 0 < y < \infty \tag{11.17}$$

with $\alpha = \delta/(1-\delta)$, $k_3 = (1-\delta)(k_2\delta^{\delta})^{1/(1-\delta)}$, and thus

$$\tilde{J}(y;\nu) \leqq k_1 + k_3 y^{-\alpha} e^{\alpha L} EZ_\nu^{\alpha}(T), \tag{11.18}$$

where L is the constant of Standing Assumption 2.1. But now

$$Z_\nu^{\alpha}(T) = \exp\left\{\alpha \int_0^T (\theta(t)+\nu(t))^* \, dW(t) - \frac{\alpha^2}{2}\int_0^T (\|\theta(t)\|^2 + \|\nu(t)\|^2)\, dt\right\}$$
$$\cdot \exp\left[\frac{\alpha(\alpha+1)}{2}\int_0^T (\|\theta(s)\|^2 + \|\nu(s)\|^2)\, ds\right],$$

and if we take $\nu \in K(\sigma)$ to satisfy

$$\int_0^T \|\nu(s)\|^2 \, ds \leqq C \quad \text{a.s.}$$

(by analogy with (5.3)) we obtain $EZ_\nu^{\alpha}(T) \leqq e^{\alpha(1+\alpha)C}$. Back into (11.18), this estimate shows that (11.3) is satisfied.

12. Existence in the dual problem. We will establish here the following existence result for the dual optimization problem of (11.2). This theorem is the final step in the solution to the problem of optimal investment in an incomplete market; see Theorem 12.5.

Assumption 12.1. $E\int_0^T \|\theta(t)\|^2 \, dt < \infty$.

Assumption 12.2. $x \mapsto xU'(x)$ is nondecreasing on $(0,\infty)$.

THEOREM 12.3. *Suppose that Assumptions* 4.3, 11.1, 11.2, 12.1, *and* 12.2 *hold. Then for every* $y \in (0,\infty)$, *there exists a process* $\lambda_y \in K_1(\sigma)$ *which achieves the infimum in* (11.2). □

Remark 12.4. Assumption 12.2 is equivalent to

$$y \mapsto yI(y) \quad \text{is nonincreasing on } (0,\infty). \tag{12.1}$$

If U is of class $C^2(0,\infty)$, Assumption 12.2 amounts to the statement that the Arrow–Pratt measure of relative risk aversion does not exceed one:

$$-\frac{xU''(x)}{U'(x)} \leqq 1 \quad \forall x \in (0,\infty). \tag{12.2}$$

On the other hand, it follows from Assumption 12.2 that $U'(x) \geqq U'(1)/x$ for all $x \geqq 1$, whence $U(x) \geqq U(1) + U'(1)\log x$. Consequently,

$$U(\infty) = \infty. \tag{12.3}$$

From this remark and Theorems 12.3, 11.8, and 9.4, we deduce then the fundamental result of §§ 11 and 12 as follows.

THEOREM 12.5. *Under the assumptions of Theorem* 12.3, *corresponding to every* $x > 0$, *there exist*:

(i) *An optimal portfolio* $\hat{\pi}_x \in \mathscr{A}(x)$ *for the utility maximization problem of* § 5; *and*

(ii) *A process* $\lambda \in K_1(\sigma)$ (λ *depending on* x) *which achieves the infimum in* (11.2) *with* $y = \mathscr{Y}_\lambda(x)$; *this* λ *satisfies the equivalent conditions* (B)-(E) *of* § 9, *and the* $\hat{\pi}$ *appearing in* (B) *is* $\hat{\pi}_x$.

We carry out the proof of Theorem 12.3 in a series of lemmas that take up much of the remainder of the section. Let us start with a rather simple observation.

LEMMA 12.6. *Under Assumptions* 11.1 *and* 12.2, *we have* $\tilde{U}(0)=\infty$, $\tilde{U}(\infty)>-\infty$, *and*

$$z\mapsto\tilde{U}(e^z)\quad \textit{is convex on } \mathbb{R}. \tag{12.4}$$

Proof. The first two claims follow directly from Lemma 4.2 and (12.3). As for (12.4), observe from (4.4) and (12.1) that $(d/dz)\tilde{U}(e^z)=-e^zI(e^z)$ is a nondecreasing function of z. □

Introduce now the Hilbert space

$$H(\sigma)\triangleq\left\{\nu\in K(\sigma);\, E\int_0^T\|\nu(s)\|^2\,ds<\infty\right\} \tag{12.5}$$

with inner product $\langle\mu,\nu\rangle\triangleq E\int_0^T\mu^*(s)\nu(s)\,ds$ and norm $[\nu]\triangleq\sqrt{\langle\nu,\nu\rangle}$. For a fixed $y>0$, we consider the functional $\tilde{J}_y: H(\sigma)\to\mathbb{R}\cup\{+\infty\}$ given by (11.2), namely,

$$\tilde{J}_y(\nu)\triangleq E\tilde{U}(y\beta(T)\,e^{-\zeta_\nu(T)}) \tag{12.6}$$

with the notation

$$\zeta_\nu(t)\triangleq\int_0^t(\theta(s)+\nu(s))^*\,dW(s)+\frac{1}{2}\int_0^t(\|\theta(s)\|^2+\|\nu(s)\|^2)\,ds,\qquad \nu\in K(\sigma). \tag{12.7}$$

LEMMA 12.7. *Under Assumptions* 11.1, 12.1, *and* 12.2, $\tilde{J}_y(\cdot)$ *is a convex functional on* $H(\sigma)$, *which satisfies*

$$\lim_{[\nu]\to\infty}\tilde{J}_y(\nu)=\infty, \tag{12.8}$$

for every $y\in(0,\infty)$.

Proof. From the convexity of the Euclidean norm in $\mathbb{R}^d$, the decrease of $\tilde{U}$ and (12.4), we have

$$\begin{aligned}\tilde{J}_y(\lambda_1\nu_1+\lambda_2\nu_2)&\leqq E\tilde{U}\left(\exp\left[\log y-\int_0^T r(s)\,ds-\lambda_1\zeta_{\nu_1}(T)-\lambda_2\zeta_{\nu_2}(T)\right]\right)\\&=E\tilde{U}\left(\exp\left\{\lambda_1\left[\log y-\int_0^T r(s)\,ds-\zeta_{\nu_1}(T)\right]\right.\right.\\&\qquad\left.\left.+\lambda_2\left[\log y-\int_0^T r(s)\,ds-\zeta_{\nu_2}(T)\right]\right\}\right)\\&\leqq\lambda_1\tilde{J}_y(\nu_1)+\lambda_2\tilde{J}_y(\nu_2)\end{aligned} \tag{12.9}$$

for any ν_1, ν_2 in $H(\sigma)$ and $\lambda_1\geqq 0$, $\lambda_2\geqq 0$ with $\lambda_1+\lambda_2=1$. On the other hand, with L as in Standing Assumption 2.1, we obtain from (12.4) and Jensen's inequality

$$\begin{aligned}\tilde{J}_y(\nu)&\geqq E\tilde{U}(\exp[\log y+L-\zeta_\nu(T)])\\&\geqq\tilde{U}(\exp[\log y+L-E\zeta_\nu(T)])\\&=\tilde{U}(\exp[\log y+L-\tfrac{1}{2}[\theta]^2-\tfrac{1}{2}[\nu]^2)\xrightarrow[{[\nu]\to\infty}]{}\infty.\end{aligned} \tag{12.10}$$

□

LEMMA 12.8. *Under Assumptions* 11.1, 12.1, *and* 12.2, *we have* $\tilde{J}_y(\nu)=\infty$, *for every* $\nu\in K(\sigma)\backslash H(\sigma)$ *and* $y\in(0,\infty)$.

Proof. Fix $y\in(0,\infty)$, $\nu\in K(\sigma)\backslash H(\sigma)$ and define stopping times

$$\tau_n \triangleq T\wedge \inf\left\{t\in[0,T];\ \int_0^t \|\nu(s)\|^2\,ds=n\right\}$$

for $n=1,2,\cdots$. With L as in Standing Assumption 2.1, it follows from Jensen's inequality, the supermartingale property of Z_ν, and the decrease of $\tilde{U}$, that

$$\begin{aligned}\tilde{J}_y(\nu)&=E\tilde{U}(y\beta(T)Z_\nu(T))=E[E\{\tilde{U}(y\beta(T)Z_\nu(T))|\mathcal{F}_{\tau_n}\}]\\&\geqq E[\tilde{U}(ye^LE\{Z_\nu(T)|\mathcal{F}_{\tau_n}\})]\geqq E\tilde{U}(ye^LZ_\nu(\tau_n))\\&\geqq \tilde{U}\left(y\exp\left\{L-\frac{1}{2}E\int_0^{\tau_n}(\|\theta(s)\|^2+\|\nu(s)\|^2)\,ds\right\}\right),\end{aligned}$$

for every $n\geqq 1$. The conclusion follows by letting $n\to\infty$. □

Proof of Theorem 12.3. Fix $y\in(0,\infty)$. The convex functional $\tilde{J}_y(\cdot)$ of (12.6) is lower-semicontinuous in the strong topology of $H(\sigma)$, by Fatou's Lemma. Therefore, it is also lower-semicontinuous in the weak topology (Ekeland and Temam [5, Cor. 2.2, Chap. 1]). Thanks to the coercivity property (12.8) of Lemma 12.7, $\tilde{J}_y(\cdot)$ attains its infimum over $H(\sigma)$ at some $\lambda_y\in H(\sigma)$ [5, Prop. 1.2, Chap. 2]. In light of Lemma 12.8 and Assumption 11.2,

$$\inf_{y\in K(\sigma)}\tilde{J}_y(\nu)=\tilde{J}_y(\lambda_y)<\infty. \tag{12.11}$$

It remains to show that $\lambda_y\in K_1(\sigma)$. From the decrease of $\tilde{U}$ and I, (4.4) and (4.8)$'$, we obtain for some $\alpha\in(0,1)$, $\gamma\in(1,\infty)$:

$$\begin{aligned}\tilde{U}(\xi)-\tilde{U}(\infty)&\geqq\tilde{U}(\xi)-\tilde{U}\left(\frac{\xi}{\alpha}\right)=\int_\xi^{\xi/\alpha}I(u)\,du\geqq\xi\left(\frac{1}{\alpha}-1\right)I\left(\frac{\xi}{\alpha}\right)\\&\geqq\frac{1-\alpha}{\alpha\gamma}\,\xi I(\xi)\quad\forall\xi\in(0,\infty).\end{aligned}\tag{12.12}$$

Replacing ξ by $y\beta(T)Z_{\lambda_y}(T)$ in (12.12) and taking expectations, we obtain

$$\mathcal{X}_{\lambda_y}(y)=E[\beta(T)Z_{\lambda_y}(T)I(y\beta(T)Z_{\lambda_y}(T))]\leqq\frac{\alpha y}{y(1-\alpha)}[E\tilde{U}(y\beta(T)Z_{\lambda_y}(T))-\tilde{U}(\infty)]<\infty,$$

thanks to (12.11), and thus $\lambda_y\in K_1(\sigma)$ by Remark 9.1. □

13. Appendix. We provide in this section an example with a well-behaved utility function U for which we do not have $\mathcal{X}_\nu(y)<\infty$ for all $y\in(0,\infty)$ and $\nu\in K(\sigma)$. The existence of such an example necessitates the introduction of the set $K_1(\sigma)$ in § 9.

In the setting of § 2, take $m=1$, $d=2$, $\sigma(t)\equiv(0,1)$, $r(t)\equiv 0$, $b(t)\equiv 0$, $T=1$, $B=W_1$, and define the stopping time $\tau\triangleq\inf\{t\in[0,1];\ t+B^2(t)=1\}$ and the process

$$\varphi(t)\triangleq\left\{\begin{matrix}\dfrac{-2B(t)}{(1-t)^2}1_{\{t\leqq\tau\}} & 0\leqq t<1\\ 0; & t=1\end{matrix}\right\},\tag{13.1}$$

$\nu(t) \triangleq (\varphi(t), 0)^*$. For the utility function $U(x) = 2\sqrt{x}$ we have $I(y) = y^{-2}$, and (7.6), (7.7) give

$$Z_\nu(t) = \exp\left\{ -\int_0^t \varphi(s)\, dB(s) - \frac{1}{2}\int_0^t \varphi^2(s)\, ds \right\}, \tag{13.2}$$

$$\mathscr{X}_\nu(y) = y^{-2} E\left[\exp\left\{ \int_0^1 \varphi(s)\, dB(s) + \frac{1}{2}\int_0^1 \varphi^2(s)\, ds \right\} \right]. \tag{13.3}$$

It is shown in Liptser and Shiryaev [15, p. 224] (see also Karatzas and Shreve [14, p. 201]) that the process Z_ν of (13.2) is *not* a martingale; in fact, the construction (13.1) is made with this property in mind. This implies, in particular, that

$$E\left[\exp\left\{ \frac{1}{2}\int_0^1 \varphi^2(s)\, ds \right\} \right] = \infty; \tag{13.4}$$

for otherwise Z_ν would be a martingale, by Novikov's theorem (Karatzas and Shreve [14, p. 199]). According to Liptser and Shiryaev [14, p. 225]:

$$\int_0^1 \varphi(s)\, dB(s) - \frac{1}{2}\int_0^1 \varphi^2(s)\, ds = -1 - 2\int_0^\tau [(1-t)^{-4} - (1-t)^{-3}] B^2(t)\, dt,$$

whence

$$\mathscr{X}_\nu(y) = \frac{1}{ey^2} E\left[\exp\left\{ 2\int_0^\tau [(1-t)^{-4} + (1-t)^{-3}] B^2(t)\, dt \right\} \right].$$

If this last expectation were finite, then so would be

$$E\left[\exp\left\{ \frac{1}{2}\int_0^1 \varphi^2(s)\, ds \right\} \right] = E\left[\exp\left\{ 2\int_0^\tau (1-t)^{-4} B^2(t)\, dt \right\} \right],$$

contradicting (13.4).

Acknowledgment. We are grateful to an anonymous referee for carefully reading this paper and offering helpful advice on exposition.

Note added in proof. It has recently been discovered that, under quite general conditions, the local martingale Z_λ is actually a martingale, where λ is the process whose existence is proved in Theorem 12.5. This result is reported in [23].

REFERENCES

[1] J. M. BISMUT, *Conjugate convex function in optimal stochastic control*, J. Math. Anal. Appl., 44 (1973), pp. 384-404.

[2] J. COX AND C. HUANG, *A variational problem arising in financial economics*, MIT mimeo, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, 1986.

[3] ———, *Optimal consumption and portfolio policies when asset prices follow a diffusion process*, J. Econom. Theory, 49 (1989), pp. 33-83.

[4] D. DUFFIE AND M. JACKSON, *Optimal hedging and equilibrium in a dynamic futures market*, J. Econom. Dynamics Control, 14 (1990), pp. 21-33.

[5] J. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, and American Elsevier, New York, 1976.

[6] J. HARRISON AND D. KREPS, *Martingales and multiperiod securities markets*, J. Econom. Theory, 20 (1979), pp. 381-408.

[7] J. M. HARRISON AND S. PLISKA, *Martingales and stochastic integrals in the theory of continuous trading*, Stochastic Process. Appl., 11 (1981), pp. 215-260.

[8] ———, *A stochastic calculus model of continuous trading: complete markets*, Stochastic Process. Appl., 15 (1983), pp. 313-316.

[9] H. HE AND N. PEARSON, *Consumption and portfolio policies with incomplete markets and short-sale constraints: the finite dimensional case*, working paper, Simon Graduate School of Business Administration, University of Rochester, Rochester, NY, 1988.

[10] ———, *Consumption and portfolio policies with incomplete markets and short-sale constraints: the infinite dimensional case*, working paper, University of California, Berkeley, CA, 1989.

[11] C. HUANG AND H. PAGÈS, *Arbitrage and optimal policies with an infinite horizon*, MIT mimeo, Massachusetts Institute of Technology, Cambridge, MA, 1989.

[12] I. KARATZAS, *Optimization problems in the theory of continuous trading*, SIAM J. Control Optim., 27 (1989), pp. 1221-1259.

[13] I. KARATZAS, J. LEHOCZKY, AND S. SHREVE, *Optimal portfolio and consumption decisions for a "small investor" on a finite horizon*, SIAM J. Control Optim., 25 (1987), pp. 1557-1586.

[14] I. KARATZAS AND S. SHREVE, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, 1988.

[15] R. S. LIPTSER AND A. N. SHIRYAEV, *Statistics of Random Processes* I: *General Theory*, Springer-Verlag, New York, 1977.

[16] R. C. MERTON, *Lifetime portfolio selection under uncertainty: the continuous-time case*, Rev. Econom. Statist., 51 (1969), pp. 247-257.

[17] ———, *Optimum consumption and portfolio rules in a continuous-time model*, J. Econom. Theory, 3 (1971), pp. 373-413. Erratum, J. Econom. Theory, 6 (1973), pp. 213-214.

[18] H. PAGÈS, *Optimal consumption and portfolio policies when markets are incomplete*, MIT mimeo, Massachusetts Institute of Technology, Cambridge, MA, 1987.

[19] S. PLISKA, *A stochastic calculus model of continuous trading: optimal portfolios*, Math. Oper. Res., 11 (1986), pp. 371-382.

[20] P. A. SAMUELSON, *Lifetime portfolio selection by dynamic stochastic programming*, Rev. Econom. Statist., 51 (1969), pp. 239-246.

[21] L. E. O. SVENSSON, *Portfolio choice and asset pricing with nontraded assets*, Institute for International Economic Studies, University of Stockholm, Stockholm, Sweden, 1988.

[22] G.-L. XU, *Ph.D. dissertation*, Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA, 1990.

[23] I. KARATZAS, J. LEHOCZKY, AND S. SHREVE, *Equivalent martingale measures and optimal market completions*, submitted, 1990.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 3, pp.731–750, May 1991

012

INVARIANCE OF THE APPROXIMATELY REACHABLE SET UNDER NONLINEAR PERTURBATIONS *

KOICHIRO NAITO † AND THOMAS I. SEIDMAN ‡

Abstract. This paper considers nonlinear perturbations of control systems with linear dynamics and seeks to analyze whether the approximately reachable set may be left unchanged by this perturbation. Under suitable conditions it is shown that this analysis may be reduced to the presumably simpler analysis of such invariance for a family of affine perturbations. Interest centers on the context of infinite-dimensional state spaces so the system may, for example, correspond to a hyperbolic or parabolic partial differential equation.

Key words. approximate reachability, control, distributed parameter system, nonlinear, perturbation, invariance

AMS(MOS) subject classifications. 49E25, 35B37, 93B05, 93C20

1. Introduction.

Reachability results for distributed parameter systems are hard to come by since, when the state space $\mathcal{X}$ is infinite-dimensional, the "standard" approaches (typically based on such implicit assumptions as local compactness, etc.) may no longer be applicable. This is especially true for nonlinear systems in a setting for which the information available relates only to the *approximately* reachable sets for related linear problems. Our concern here is with the extension to this more general context of a mode of analysis treated in a sequence of earlier papers [11], [12], [15], [16].

We consider a nonlinear control system given by

$$\dot{x} = \mathbf{A}x + \mathbf{F}x + \mathbf{B}u, \qquad u \in \boldsymbol{\mathcal{U}}_{ad} \tag{1.1}$$

with $\mathbf{A}$ linear and $\mathbf{F}$ a nonlinear operator satisfying some suitable growth condition. The control operator $\mathbf{B}$ need not be linear and we may occasionally write $\mathbf{B}(u)$ to emphasize this; we set $\boldsymbol{\mathcal{W}}_{ad} := \{w = \mathbf{B}(u) : u \in \boldsymbol{\mathcal{U}}_{ad}\}$.

We wish to consider (1.1) as a "perturbation" of the control problem

$$\dot{z} = \mathbf{A}z + \mathbf{B}u, \qquad u \in \boldsymbol{\mathcal{U}}_{ad} \tag{1.2}$$

omitting the nonlinearity $\mathbf{F}$. For the moment we require only that (1.2) should have a meaningful solution for each $u \in \boldsymbol{\mathcal{U}}_{ad}$ and we will consistently denote this solution by z_u. Setting

$$\boldsymbol{\mathcal{Z}}_{ad} := \{z_u \text{ satisfying (1.2) for some } u \in \boldsymbol{\mathcal{U}}_{ad}\}, \tag{1.3}$$

part of our interpretation of "meaningful" is that $\boldsymbol{\mathcal{Z}}_{ad}$ is to be considered in some[1] Banach space $\boldsymbol{\mathcal{X}}$ of functions on $[0,T]$ with continuity to $\mathcal{X}$ for the evaluation map

* Received by the editors December 29, 1986; accepted for publication (in revised form) March 17, 1990.

† Department of Economics, Senshuu University, Higashimita, Tama-ku, Kawasaki-shi, 152 Japan (e-mail: knaito@ponta.sucis.senshu-u.ac.jp).

‡ Department of Mathematics and Statistics, University of Maryland Baltimore County, Baltimore, Maryland 21228 (e-mail: seidman@umbc.bitnet, @umbc1.umbc.edu). This research has been partially supported by the United States Air Force Office of Scientific Research under grants AFOSR-82-0271 and AFOSR-87-0350 and by the National Science Foundation under grants CDR 85-00108 and ECS-88-14788. Acknowledgment is also due to the System Research Center (University of Maryland) for its hospitality.

[1] In general we take $\boldsymbol{\mathcal{X}}$ to be $C([0,T] \to \mathcal{X})$, which certainly ensures that the evaluation map $\mathbf{E}$ will be continuous. We do note that lesser regularity could be permitted away from the terminal time — in particular, to allow for the effect of "rough" initial data.

$\mathbf{E} : z \mapsto z(T) : \boldsymbol{\mathcal{X}} \to \mathcal{X}$ so that we can speak about the *reachable set*

$$\mathcal{K}_0 := \mathbf{E}\boldsymbol{\mathcal{Z}}_{ad} := \{\mathbf{E}z : z \in \boldsymbol{\mathcal{Z}}_{ad}\}$$

as a subset of the state space $\mathcal{X}$.

In considering (1.1), we observe that: for any *particular* solution $x = x_u$, the nonlinearity gives $\mathbf{F}x$ as some specific function g. This solution is then also a solution of an alternate perturbation of the system (1.2):

$$(1.4) \qquad \dot{x} = \mathbf{A}x + g + \mathbf{B}u, \qquad u \in \boldsymbol{\mathcal{U}}_{ad}$$

using the same control u. We will consider the family of *affine perturbations* (1.4) with g taken from some given subset $\boldsymbol{\mathcal{G}}$ of a function space $\boldsymbol{\mathcal{V}}$. The relation, of course, is that $\boldsymbol{\mathcal{G}}$ is to contain each $\mathbf{F}[x(\cdot)]$ as u ranges over $\boldsymbol{\mathcal{U}}_{ad}$ in (1.1). Presumably the reachability analysis for (each of) the affinely perturbed problems (1.4) should be simpler than that for the original nonlinearly perturbed problem (1.1). For (1.4), (1.1) we will again be considering solutions in the same space $\boldsymbol{\mathcal{X}}$ of $\mathcal{X}$-valued functions as we use for the "unperturbed" equation (1.2).

It will be important for us to consider the relation of (1.1) to (1.2) from the viewpoint of (1.4). Given a solution $z = z_u$ for (1.2), we may write (1.1) as an equation for $y = x - z$:

$$(1.5) \qquad \mathbf{G} : z \mapsto \mathbf{F}(y + z) =: g \ \text{ such that } \dot{y} = \mathbf{A}y + g.$$

For the moment this specification of a "map" $\mathbf{G} = \mathbf{G}_{\mathrm{F}}$ is purely formal but, of course, we will eventually want to have $\mathbf{G} : \boldsymbol{\mathcal{Z}}_{ad} \to \boldsymbol{\mathcal{G}}$.

The question we wish to consider is:

(Q) When must a state ξ which is approximately reachable for (1.2) also be approximately reachable for the nonlinear system (1.1)?

Letting $\mathcal{K}_{\mathrm{F}} \subset \mathcal{X}$ denote the reachable set for (1.1) — i.e., the set of values for $\xi = x(T)$ as x ranges over solutions of (1.1) with $u \in \boldsymbol{\mathcal{U}}_{ad}$ — this asks whether $\mathcal{K}_0 \subset \bar{\mathcal{K}}_{\mathrm{F}}$; we note that the complementary inclusion is, in general, much easier to analyze. More precisely, we are asking:

> When can the question (**Q**) above be analyzed (answered affirmatively) by considering the *family* of (presumably simpler) related questions for the affine problems (1.4) as g ranges over $\boldsymbol{\mathcal{G}}$?

In this form it should be clear that our concern is with the *validation of a mode of analysis.* We are, for example, seeking conditions on (1.4) for $g \in \boldsymbol{\mathcal{G}}$ ensuring that $\bar{\mathcal{K}}$ is invariant under the perturbation: (1.2) $\longmapsto$ (1.1). The essential requirement will be a certain uniformity (over suitable bounded subsets of $\boldsymbol{\mathcal{G}}$) for a measure of the approximate reachability; see (3.23) et seq.

We have already investigated the corresponding invariance of the (exactly) reachable set in a sequence of papers [11], [12], [15], [16]. The work presented here represents an extension of this work in two directions: the consideration of *approximate* rather than exact reachability and *the consideration of control sets which do not form a linear space with linear* **B**.

We further note that our present approach to (**Q**) is "one ξ at a time" and so, with minor modification, lends itself to an analysis of approximate controllability for (1.1) in a less "global" context than the invariance of the title. The arguments used here put this work in the setting of "the fixed point approach to controllability" and we refer, e.g., to [5] and its references for further historical discussion of this approach.

2. Some examples. In this section, before beginning the analysis of reachability, we wish to provide some representative examples for which one can verify certain general hypotheses, $(\mathbf{H}_1)$ below, as to the settings to which our analysis applies. As previously noted, our concern is with settings for which the state space $\mathcal{X}$ is infinite-dimensional so that (1.2) may itself be an abstract formulation of a partial differential equation. We emphasize that our hypotheses do *not* restrict consideration to parabolic equations (compact, analytic semigroups). The analysis here could, in principle, be applied also to hyperbolic systems — although one expects that considerations of *approximate* reachability (compare [15], [17]) might then only be of interest if $\{\mathbf{B}(u) : u \in \boldsymbol{\mathcal{U}}_{ad}\}$ is not itself a linear space.

We turn now to some specific examples of systems governed by partial differential equations whose dynamics will satisfy the general hypotheses $(\mathbf{H}_1)$ below — but only indicate briefly the verifications. In §6 we will provide, following [15], some results which are sufficient to provide such verification for more general classes of problems. For each of these examples the spatial domain will be a bounded region $\Omega \subset \mathbb{R}^m$ with "smooth enough" boundary $\partial\Omega$; we set $\mathcal{Q} := (0,T) \times \Omega$ and $\Sigma := (0,T) \times \partial\Omega$.

Example 1. Our first example already exhibits many of our concerns for application. Consider boundary control of the nonlinear heat equation:

$$\dot{x} = \Delta x + \varphi(\cdot, x, \nabla x) \text{ on } \mathcal{Q}, \qquad x_n = e^u \text{ on } \Sigma \tag{2.6}$$

as (1.1). The corresponding "linear" control problem is then

$$\dot{z} = \Delta z \text{ on } \mathcal{Q}, \qquad z_n = w \text{ on } \Sigma \tag{2.7}$$

where $w := e^u$. (Note that x_n, z_n denote normal derivatives at $\partial\Omega$.)

We wish to take $\mathcal{X} := L^2(\Omega)$ as state space. Note that the function $f(\xi) := \varphi(\cdot, \xi(\cdot), [\nabla\xi](\cdot))$ is not well defined for general $\xi \in \mathcal{X}$ since the argument $\nabla\xi$ would then only be in $H^{-1}(\Omega)$ and so need not be a function at all for pointwise composition with φ. This is a situation for which the considerations of Lemma 3 below are important. If we assume that the scalar function φ is (uniformly) Lipschitzian in its variables $\xi, \nabla\xi$, then we can introduce the auxiliary space $\boldsymbol{\mathcal{Y}} := L^2([0,T] \to H^1(\Omega))$ and have Lipschitz continuity from $\boldsymbol{\mathcal{Y}}$ to $\boldsymbol{\mathcal{V}} := L^2(\mathcal{Q})$ for the Nemytsky operator $\mathbf{F} : x \mapsto \varphi(\cdot, x, \nabla x)$. Fairly standard methods show that $\mathbf{L}$ is well defined and continuous from this $\boldsymbol{\mathcal{V}}$ to $\boldsymbol{\mathcal{X}}$ and to $\boldsymbol{\mathcal{Y}}$. Further, if one uses a suitable weighting (exponential in t) for the norms of $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{V}}$, then one can arrange that $\vartheta < 1$ so Lemma 3 will apply (provided we can take $\boldsymbol{\mathcal{Z}} \hookrightarrow \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{Y}}$).

Note that these considerations do not yet involve $\boldsymbol{\mathcal{U}}_{ad}$ or $\boldsymbol{\mathcal{Z}}_{ad}$ as $\mathbf{L}$ relates only to the problem with *homogeneous* boundary conditions (and, of course, homogeneous initial conditions).

Suppose we would wish to consider $\boldsymbol{\mathcal{U}}_{ad} = \{u \in L^2(\Sigma) : u > 0 \text{ ae}\}$. Clearly we must begin by restricting this further, to those u (and corresponding $w = e^u$) for which (2.7) and (2.6) have solutions in some satisfactory sense. If we take $w \in \boldsymbol{\mathcal{W}} := L^2(\Sigma)$, then it is standard that the solution map $\mathbf{S}$ defined by (2.7) is continuous and, indeed, compact from $\boldsymbol{\mathcal{W}}$ to $\boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{Y}}$. One can then take, e.g., $\boldsymbol{\mathcal{Z}} := \mathbf{S}\boldsymbol{\mathcal{W}}$, normed so that $\mathbf{S}$ is an isometry. By suitable density arguments one can show that the $\boldsymbol{\mathcal{X}}$-closures of the sets of solutions for (2.6), (2.7), and the corresponding equation affinely perturbed by arbitrary $g \in \boldsymbol{\mathcal{V}}$ will be the same whether one works with $\boldsymbol{\mathcal{W}}_{ad} := \{w \in \boldsymbol{\mathcal{W}} : w \geq 1 \text{ ae}\}$ or with the original $\boldsymbol{\mathcal{U}}_{ad}$, restricted to u giving solutions in $\boldsymbol{\mathcal{X}}$. Thus one has equivalence in the sense of (3.17), below. This, together with Lemma 3, gives $(\mathbf{H}_1)$.

We note also that this $\boldsymbol{\mathcal{Z}}_{ad}$ is convex and closed in $\boldsymbol{\mathcal{Z}}$ since we obviously have $\boldsymbol{\mathcal{W}}_{ad}$ convex and closed in $\boldsymbol{\mathcal{W}}$.

Example 2. We next consider distributed control of the equation

$$\dot{x} = \Delta x + \varphi(\cdot, x, \nabla x) + w \text{ on } \mathcal{Q}, \qquad x = 0 \text{ on } \Sigma, \tag{2.8}$$

assuming that $\varphi : \mathcal{Q} \times \mathbb{R}^{m+1} \to \mathbb{R}$ is a "nice" function (i.e., with uniform bounds on enough derivatives). Again, (2.8) can be put into the abstract form (3.15) by taking the state space $\mathcal{X}$ to be $\mathcal{X}_0 = L^2(\Omega)$ and then introducing $\mathbf{A}$ as the Laplacian on Ω with domain $\mathcal{D}(\mathbf{A}) := \{\xi \in H^2(\Omega) \subset \mathcal{X}_0 : \xi|_{\partial\Omega} = 0\}$. Since $-\mathbf{A}$ is self-adjoint and positive, there is no difficulty (e.g., through the rate of decay of the eigenfunction expansion coefficients) in defining $[-\mathbf{A}]^r$ for all $r \geq 0$ and we set $\mathcal{X}_\mu := \mathcal{D}([-\mathbf{A}]^{\mu/2})$; it is known [7], [8] that $\mathcal{X}_\mu = H^\mu(\Omega)$ for $0 \leq \mu < \frac{1}{2}$, that $\mathcal{X}_\mu = \{\xi \in H^\mu(\Omega) : \xi|_{\partial\Omega} = 0\}$ for $\frac{1}{2} < \mu < \frac{5}{2}$, etc.

Consider controls $w \in \boldsymbol{\mathcal{W}} = \boldsymbol{\mathcal{W}}_m := L^2([0,T] \to \mathcal{X}_m)$ for (fixed) $m \geq 0$. A standard semigroup estimate[2] gives continuity for $\mathbf{L}_t$ from $\boldsymbol{\mathcal{W}}_m$ to $\mathcal{X}_\mu$ for each $\mu < m+1$ and a convolution estimate shows that $\mathbf{L}w =: x(\cdot)$ will be in $L^2([0,T] \to \mathcal{X}_\mu$ for $\mu < m+2$. We get

$$x(\cdot) \in \boldsymbol{\mathcal{X}}_\mu := C([0,T] \to \mathcal{X}_\mu) \cap L^2([0,T] \to \mathcal{X}_{\mu+1})$$

for each $\mu < m+1$. On the other hand, for any $\xi \in \mathcal{X}_{m+2}$ we can set $w_*(t) := (\xi - t\mathbf{A}\xi)/T$ to get $w_* \in \boldsymbol{\mathcal{W}}_m$ and $\mathbf{L}_t w_* = (t/T)\xi$. Hence, $\mathcal{X}_{m+2} \subset \mathcal{K}_0(\boldsymbol{\mathcal{W}}_m) \subset \cap_{\mu<m+1}\mathcal{X}_m$. Any *more* precise estimate of $\mathcal{K}_0(\boldsymbol{\mathcal{W}}_m)$ would be much more difficult to obtain — even in this "simple" setting.

In considering the nonlinear partial differential equation (2.8), our assumptions make φ uniformly Lipschitzian in $x, \nabla x$, so a standard (Picard iteration) argument gives existence of a (unique) solution of the nonlinear equation (2.8) in $\boldsymbol{\mathcal{X}}_0$ for, say, any control $w \in \boldsymbol{\mathcal{W}}_0 = L^2(\mathcal{Q})$; compare the proof of Lemma 3 below and the proof of Lemma 4 in [15]. This then gives $\varphi(\cdot, x, \nabla x) =: \mathbf{F}x =: \mathbf{G}_\mathbf{F} w =: g$ again in $L^2(\mathcal{Q})$ so $[g + w] \in \boldsymbol{\mathcal{W}}_0$ and by the analysis above, we then have $x \in \boldsymbol{\mathcal{X}}_\mu$ for arbitrary $\mu < 1$ and this gives $\mathcal{K}_\mathbf{F} \subset \cap_{\mu<1}\mathcal{X}_\mu$. For smoother controls one can similarly obtain the regularity result: $\mathcal{K}_\mathbf{F}(\boldsymbol{\mathcal{W}}_m) \subset \cap_{\mu<m+1}\mathcal{X}_\mu$ provided one has an estimate of the form:

$$\|(-\mathbf{A})^{\mu/2}\mathbf{F}x\| \leq C[1 + \|(-\mathbf{A})^{(\mu+1)/2}x\|] \tag{2.10}$$

for such μ. The estimate (2.10) follows from the assumed regularity of φ while $\mu < \frac{1}{2}$ but the boundary conditions then intervene; one will only continue to have (2.10), at least while $\mu < \frac{5}{2}$, provided that, in addition to the bounds on derivatives for φ, one were to require that $\varphi(\cdot, 0, \cdot) = 0$ on $\partial\Omega$ so $x|_{\partial\Omega} = 0$ also implies $[\mathbf{F}x]|_{\partial\Omega} = 0$.

For the nonlinear equation (2.8) there is no longer a simple explicit calculation to give a useful *lower* bound for $\bar{\mathcal{K}}_\mathbf{F}(\boldsymbol{\mathcal{W}}_m)$. Taking $\mathcal{X} = \mathcal{X}_0$, for example, we have $\mathcal{X}_{m+2}$ dense in $\mathcal{X}$ for arbitrary m so $\bar{\mathcal{K}}_0(\boldsymbol{\mathcal{W}}_m) = \mathcal{X}$ and we would similarly like to show that $\bar{\mathcal{K}}_\mathbf{F}(\boldsymbol{\mathcal{W}}_m) = \mathcal{X}$. Under the assumptions we are now imposing, we can consider, e.g., $1 < m < \frac{3}{2}$ and get

$$\mathbf{L}_{[T]}g \in \mathcal{D}(\mathbf{A}) = [\mathcal{D}([-\mathbf{A}]^{m+1}), \mathcal{X}]_\vartheta \subset [\mathcal{K}_0(\boldsymbol{\mathcal{W}}_m, \mathcal{X}]_\vartheta$$

[2] Since $\mathbf{A}$ generates an analytic semigroup on $\mathcal{X}_0$ one has

$$\|(-\mathbf{A})^r\mathbf{S}(\tau)\| \leq C\tau^{-r} \tag{2.9}$$

on bounded intervals for $r \geq 0$ and with C depending on r and the interval.

with $\vartheta = m/(m+1)$ as parameter for the interpolation spaces, corresponding to (5.41). Thus, looking ahead, we note that the assumptions of Theorem 5 will then be satisfied if one further imposes a growth condition on φ to strengthen (2.10) so as to get (6.48) with $r < 1/(m+1)$.

Example 3. For our next example we take $\Omega = (0,1) \subset \mathbb{R}^1$ and consider

$$\begin{aligned} &\dot{x} = x'' + \varphi(x, x') &&\text{on} \quad \mathcal{Q} := (0,T) \times (0,1), \\ &x(\cdot, 0) = u(\cdot), \ x(\cdot, 1) = 0 &&\text{on} \quad (0,T). \end{aligned} \tag{2.11}$$

As above we assume φ is uniformly Lipschitzian and take $\mathcal{X} := L^2(0,1)$; we now wish to take

$$\boldsymbol{\mathcal{W}}_{ad} = \boldsymbol{\mathcal{U}}_{ad} := \{\text{integer-valued step functions on } (0,T)\}.$$

We again take $\boldsymbol{\mathcal{Y}} := L^2([0,T] \to H^1(0,1))$ and $\boldsymbol{\mathcal{V}} := L^2(\mathcal{Q})$. Since we can "restart" the equations at each of the (finitely many) jumps of $u \in \boldsymbol{\mathcal{U}}_{ad}$, there is no difficulty with the solvability, for each such u, of the various equations: solutions will be piecewise smooth. Our difficulty is now with the continuity of the maps and with the desired convexity of $\boldsymbol{\mathcal{Z}}_{ad}$.

Suppose the scalar function φ also satisfies a growth condition $\varphi(r,s) \le a_0 + a_1|r|$ so the Nemytsky operator $\mathbf{F}$ satisfies $|[\mathbf{F}x](t)|_{\mathcal{X}} \le C_0 + C_1|x(t)|_{\mathcal{X}}$, independent of x'. In this case, a bound on u in, say, $L^2(0,T)$ gives solutions, for (2.11) and for the corresponding linear equations, in a compact subset of $\boldsymbol{\mathcal{X}}$ and, using interior regularity, such that the corresponding gradients are in a compact subset of, say, $L^2((0,T)\times(a,b))$ for any $0 < a < b < 1$. It follows, extracting suitable convergent subsequences for which the gradients are converging pointwise ae, that if one considers any bounded set in $L^2(0,T)$ then all limits of the solutions will again be solutions in $\boldsymbol{\mathcal{X}}$. We now observe (compare Lemma 5) that the set $\boldsymbol{\mathcal{U}}_{ad}$ above and the set $\boldsymbol{\mathcal{W}} := H_0^1(0,T)$ have the same (sequential) weak closure in $L^2(0,T)$, i.e., all of $L^2(0,T)$, so the $\boldsymbol{\mathcal{X}}$-closures of the sets of solutions (and so the corresponding approximately reachable sets) will be the same for each of these as for $L^2(0,T)$ — restricted to those u for which there are, indeed, solutions in $\boldsymbol{\mathcal{X}}$ satisfying the boundary conditions in a meaningful sense. It follows that we can take $\boldsymbol{\mathcal{W}}'_{ad} := \boldsymbol{\mathcal{W}}$ and $\boldsymbol{\mathcal{Z}}'_{ad} = \boldsymbol{\mathcal{Z}} := \mathbf{S}\boldsymbol{\mathcal{W}}'_{ad}$ (noting that we do have continuity and compactness for $\mathbf{S} : \boldsymbol{\mathcal{W}} \to \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{Y}}$) with equivalence in the sense of (3.17). We will have $(\mathbf{H}_1)$ and the needed convexity if we work with this $\boldsymbol{\mathcal{Z}}_{ad}$.

Example 4. For our final example we consider a quasi-linear wave equation

$$\ddot{x} = \Delta x + \varphi(\cdot, x) \text{ on } \mathcal{Q}, \qquad x|_{\Sigma} = u \tag{2.12}$$

where we assume φ is smooth, uniformly Lipschitzian and with a growth condition $|\varphi| \le C(1 + |x|)^{\bar{r}}$ with $\bar{r} < 1$. We assume homogeneous initial conditions: $x(0) = 0 = \dot{x}(0)$ and that the controls $u \in \boldsymbol{\mathcal{U}}_{ad}$ are to be taken smooth enough to have an extension to $\mathcal{Q}$, again denoted by u, with, say, $q := [\ddot{u} - \Delta u] \in L^2(\mathcal{Q})$. This will ensure that the solution $z = z_u$ of the corresponding form of (1.2)

$$\ddot{z} = \Delta z \text{ on } \mathcal{Q}, \qquad z|_{\Sigma} = u \tag{2.13}$$

will be in $\boldsymbol{\mathcal{Z}} := C^1([0,T] \to L^2(\Omega) \cap C([0,T] \to H^1(\Omega))$, using an estimate obtained by multiplying (2.13) by $[z - u]$, integrating, and applying the Gronwall Inequality. If we take $\boldsymbol{\mathcal{V}} := L^2(\mathcal{Q})$, then a similar estimate shows the continuity of $\mathbf{L} : \boldsymbol{\mathcal{V}} \to \boldsymbol{\mathcal{Z}}$. Finally, we note that (1.5) becomes

$$\ddot{y} = \Delta y + \varphi(\cdot, y + z) \text{ on } \mathcal{Q}, \qquad y|_{\Sigma} = 0$$

and a standard contraction mapping argument gives existence of a solution $y \in \boldsymbol{\mathcal{Z}}$, using the Lipschitz property of φ, with an estimate $|y| = \mathcal{O}(|z|^{\bar{r}})$ where $|y|$ is the $\boldsymbol{\mathcal{Z}}$-norm and $|z|$ is the $\boldsymbol{\mathcal{V}}$-norm. The compactness of the embedding $\boldsymbol{\mathcal{Z}} \hookrightarrow \boldsymbol{\mathcal{V}}$ then completes the verification of the general hypotheses $(\mathbf{H}_1)$ with the growth condition (4.28).

We note two particular cases of interest from the viewpoint of reachability: (1) in the one-dimensional case with $\boldsymbol{\mathcal{U}}_{ad}$ consisting of functions vanishing at one end of Ω and nonnegative at the other, it is known [13] that $\mathcal{K}_0$ is dense in, say, $\mathcal{X} = L^2(\Omega)$ for T large enough; (2) for $\Omega \subset \mathbb{R}^m$ with $u \in \boldsymbol{\mathcal{U}}_{ad}$ having support in some small fixed subset $\Gamma \subset \partial\Omega$ and/or with T not too large, then the reachable set will only be some (small) part of $\mathcal{X}$ and it is interesting to ask whether a geometric restriction on the support in $\mathcal{Q}$ for φ could provide the hypotheses for invariance.

3. Formulation and notation. We have already introduced the set $\boldsymbol{\mathcal{Z}}_{ad}$ of solutions of (1.2) (in some fixed sense) as the control u ranges over $\boldsymbol{\mathcal{U}}_{ad}$ and the (formal) operator $\mathbf{E} : x \mapsto x(T)$. We now also introduce the linear solution operator $\mathbf{L}$:

$$\mathbf{L} : v \mapsto x \text{ such that } \dot{x} = \mathbf{A}x + v \tag{3.14}$$

(with homogeneous initial conditions) for suitable $v(\cdot)$. Note that (1.2) and (1.4) have linear dynamics (i.e., linear in z or x although not necessarily in u) while the dynamics given by the perturbed equation (1.1) are quasi-linear.

We have already made our *first basic observation*: neither $\mathbf{B}$ nor $u \in \boldsymbol{\mathcal{U}}_{ad}$ (nor their individual properties) can be relevant to any of (1.2), (1.1), (1.4), but only $w := \mathbf{B}(u)$, considered as an element of some space $\boldsymbol{\mathcal{W}}$. Indeed, we have seen that the only *really* relevant entity is $z = z_u$, the corresponding solution of (1.2). Thus, the effects of control are entirely determined by the set $\boldsymbol{\mathcal{Z}}_{ad}$.

Expressed in terms of this (formal) operator $\mathbf{L}$, the differential equations (1.1) and (1.4) then take the abstract forms:

$$x = \mathbf{LF}x + z \qquad (z \in \boldsymbol{\mathcal{Z}}_{ad}), \tag{3.15}$$

$$x = \mathbf{L}g + z \qquad (z \in \boldsymbol{\mathcal{Z}}_{ad},\ g \in \boldsymbol{\mathcal{G}}), \tag{3.16}$$

where $\boldsymbol{\mathcal{Z}}_{ad}$ is now simply a (specified) subset of some function space $\boldsymbol{\mathcal{Z}}$ and $\boldsymbol{\mathcal{G}}$ is a specified subset of another function space $\boldsymbol{\mathcal{V}}$. Until one specifies the spaces involved this is purely formal but we note here that, although we refer for convenience to (1.1), (1.2), (1.4), we will always be interpreting "solution" in the present sense: through the *abstract operator equations* (3.15), (3.16) with any hypotheses and interpretations to be attached to these. Except for this section, the (motivating) earlier examples, and the final section, our considerations are independent of any interpretation of (3.15) and (3.16) as differential equations.

Note that, since we consider the equations (1.2), etc., with fixed initial conditions, it is always possible, with no loss of generality, to translate the problem by some fixed trajectory z_0, correspondingly modifying $\mathbf{F}$, $\boldsymbol{\mathcal{Z}}_{ad}$, and all reachable sets. Henceforth, for expository simplicity, we do assume that, ab initio, the problem has been formulated with homogeneous initial conditions so that one has $z_u = \mathbf{L}w$ where $w = \mathbf{B}(u)$. At the same time, once one has avoided consideration of any (essentially irrelevant) problems with regularity near the initial time, it is convenient to assume that all our solutions are elements of the fixed space $\boldsymbol{\mathcal{X}} := C([0,T] \to \mathcal{X})$ so that, in particular, the operator $\mathbf{E} : \boldsymbol{\mathcal{X}} \to \mathcal{X}$ is always well defined and continuous. We also assume that $\boldsymbol{\mathcal{Z}}_{ad} \subset \boldsymbol{\mathcal{X}}$ set-theoretically, but may find it convenient to topologize it somewhat differently: as $\boldsymbol{\mathcal{Z}}_{ad} \subset \boldsymbol{\mathcal{Z}}$; e.g., if there is a continuous embedding $\boldsymbol{\mathcal{Z}} \hookrightarrow \boldsymbol{\mathcal{X}}$.

Our underlying set of "solvability hypotheses" is:

($\mathbf{H}_1$) Let *each* of the following hold:

(i) Equation (3.14) has a (unique) solution $x = \mathbf{L}v \in \boldsymbol{\mathcal{X}}$ for each $v \in \boldsymbol{\mathcal{V}}$; the linear map $\mathbf{L}$ is well-defined and continuous from $\boldsymbol{\mathcal{V}}$ to $\boldsymbol{\mathcal{X}}$.

(ii) For each $z \in \boldsymbol{\mathcal{Z}}_{ad}$ there is a unique solution x of (3.15) and we assume that $g = \mathbf{F}x$ is in $\boldsymbol{\mathcal{V}}$, i.e., there is a well-defined (nonlinear) map $\mathbf{G} = \mathbf{G}_{\mathrm{F}} : \boldsymbol{\mathcal{Z}}_{ad} \to \boldsymbol{\mathcal{V}} : z \mapsto g$ which we assume is continuous. By ([illegible]) we then have $x \in \boldsymbol{\mathcal{X}}$ for the solution x of (3.15).

(iii) The map $\mathbf{G} : \boldsymbol{\mathcal{Z}}_{ad} \to \boldsymbol{\mathcal{V}}$ is *compact*.

Compare (1.5) for the definition of the map $\mathbf{G}$. In §4 we will consider some specific examples involving partial differential equations and some classes of settings for which these abstract hypotheses can be verified.

Having formulated the "dynamics" of the problem — introducing the relevant spaces and the operators $\mathbf{L}$ and $\mathbf{G}_{\mathrm{F}}$ to obtain (3.15), (3.16) — we next wish to consider the various reachable and approximately reachable sets. We now define $\mathbf{L}_{[t]} : v \mapsto [\mathbf{L}v](t)$ (so, in particular, $\mathbf{EL} = \mathbf{L}_{[\mathrm{T}]}$) and set

$$\mathbf{T} = \mathbf{T}_{\mathrm{F}} := \mathbf{L}_{[\mathrm{T}]}\mathbf{G} + \mathbf{E} : \boldsymbol{\mathcal{Z}}_{ad} \to \mathcal{X} : z \mapsto \xi := \mathbf{E}x \text{ such that (3.15).}$$

Clearly, in view of ($\mathbf{H}_1$), the operator $\mathbf{T}$ is continuous. If, for arbitrary $\boldsymbol{\mathcal{Z}}_* \subset \boldsymbol{\mathcal{Z}}$, we define:

$$\begin{aligned}
\mathcal{K}_{\mathrm{F}}(\boldsymbol{\mathcal{Z}}_*) &:= \{\mathbf{E}x : \text{(3.15) for } z \in \boldsymbol{\mathcal{Z}}_*\} = \{\mathbf{T}_{\mathrm{F}}z : z \in \boldsymbol{\mathcal{Z}}_*\},\\
\mathcal{K}_0(\boldsymbol{\mathcal{Z}}_*) &:= \mathbf{E}\boldsymbol{\mathcal{Z}}_* := \{\mathbf{E}z : z \in \boldsymbol{\mathcal{Z}}_*\},\\
\mathcal{K}_g(\boldsymbol{\mathcal{Z}}_*) &:= \{(\mathbf{L}_{[\mathrm{T}]}g + \mathbf{E}z) : z \in \boldsymbol{\mathcal{Z}}_*\} = \mathbf{L}_{[\mathrm{T}]}g + \mathcal{K}_0(\boldsymbol{\mathcal{Z}}_*),
\end{aligned}$$

then the (exactly) *reachable sets* for (1.2), (1.1), and (1.4) will be $\mathcal{K}_0 = \mathcal{K}_0(\boldsymbol{\mathcal{Z}}_{ad})$, $\mathcal{K}_{\mathrm{F}} = \mathcal{K}_{\mathrm{F}}(\boldsymbol{\mathcal{Z}}_{ad})$, and $\mathcal{K}_g = \mathcal{K}_g(\boldsymbol{\mathcal{Z}}_{ad})$, respectively. The corresponding approximately reachable sets are then the $\mathcal{X}$-closures: $\bar{\mathcal{K}}_0$, $\bar{\mathcal{K}}_{\mathrm{F}}$, $\bar{\mathcal{K}}_g$, respectively (i.e., $\overline{\mathcal{K}_0(\boldsymbol{\mathcal{Z}}_{ad})}$, etc.). We make here our *second basic observation*: for present purposes we may always replace the "original" $\boldsymbol{\mathcal{Z}}_{ad}$ at our convenience with any other set $\boldsymbol{\mathcal{Z}}'_{ad}$ for which the approximately reachable sets are the same:

$$\bar{\mathcal{K}}_0(\boldsymbol{\mathcal{Z}}'_{ad}) = \bar{\mathcal{K}}_0(\boldsymbol{\mathcal{Z}}_{ad}),\ \bar{\mathcal{K}}_g(\boldsymbol{\mathcal{Z}}'_{ad}) = \bar{\mathcal{K}}_g(\boldsymbol{\mathcal{Z}}_{ad}),\ \bar{\mathcal{K}}_{\mathrm{F}}(\boldsymbol{\mathcal{Z}}'_{ad}) = \bar{\mathcal{K}}_{\mathrm{F}}(\boldsymbol{\mathcal{Z}}_{ad}). \tag{3.17}$$

In particular, the continuity of $\mathbf{G}$ assumed in ($\mathbf{H}_1$)(ii) will mean that we can always take $\boldsymbol{\mathcal{Z}}_{ad}$ closed, i.e., replacing it by its closure with no loss of generality. As part of this observation we also note that $\boldsymbol{\mathcal{Z}}_{ad}$ — or, equivalently, $\boldsymbol{\mathcal{Z}}'_{ad}$ satisfying (3.17) — is of purely set-theoretic significance so the relevant topology is at our convenience.

It is sometimes convenient to introduce explicitly the intermediate space $\boldsymbol{\mathcal{W}}$ such that $\mathbf{B}(u) \in \boldsymbol{\mathcal{W}}_{ad} \subset \boldsymbol{\mathcal{W}}$ so $\boldsymbol{\mathcal{Z}}_{ad} = \mathbf{S}\boldsymbol{\mathcal{W}}_{ad}$ and to consider the map: $u \mapsto w := \mathbf{B}(u) \mapsto z = z_u$ defined by (1.2). The second "factor" of this is the (linear) solution map for (1.2):

$$\mathbf{S} : \boldsymbol{\mathcal{W}} \to \boldsymbol{\mathcal{Z}} : w \mapsto z \tag{3.18}$$

and we now write $\boldsymbol{\mathcal{W}}_{ad} \equiv \boldsymbol{\mathcal{W}}'_{ad}$ if $\boldsymbol{\mathcal{Z}}_{ad} := \mathbf{S}\boldsymbol{\mathcal{W}}_{ad}$ and $\boldsymbol{\mathcal{Z}}'_{ad} := \mathbf{S}\boldsymbol{\mathcal{W}}'_{ad}$ are equivalent in the sense of (3.17).

In our previous work we assumed, with $\boldsymbol{\mathcal{G}}_* = \boldsymbol{\mathcal{V}}$, that the *exactly* reachable set for (1.2) was invariant under all the affine perturbations (1.4) for $g \in \boldsymbol{\mathcal{V}}$, i.e., that

$$\mathcal{K}_g = \mathcal{K}_0 \text{ for each } g \in \boldsymbol{\mathcal{G}}_*. \tag{3.19}$$

The natural reachability condition for us to consider in attempting a corresponding treatment of approximate controllability would seem to be

$$\bar{\mathcal{K}}_g = \bar{\mathcal{K}}_0 \text{ for each } g \in \boldsymbol{\mathcal{G}}_* \tag{3.20}$$

but we will need to strengthen half of this, the inclusion

$$\mathcal{K}_0 \subset \bar{\mathcal{K}}_g, \tag{3.21}$$

to obtain the desired result. Suppose, given $\xi \in \mathcal{X}$ and $\varepsilon > 0$, we introduce the set-valued function $\mathcal{C}(\cdot) = \mathcal{C}(\cdot;\xi,\varepsilon)$ defined on $\boldsymbol{\mathcal{V}}$ by

$$\mathcal{C}(g;\xi,\varepsilon) := \{z \in \boldsymbol{\mathcal{Z}}_{ad} : |\xi - [\mathbf{L}_{[\mathrm{T}]}g + \mathbf{E}z]|_{\mathcal{X}} \le \varepsilon\}. \tag{3.22}$$

Then (3.21) just means that $\mathcal{C}(g;\xi,\varepsilon)$ is nonempty for each $\xi \in \mathcal{K}_0$ and each $\varepsilon > 0$. Fixing $\xi \in \mathcal{K}_0$ and $\varepsilon > 0$, we now set

$$\nu(g) = \nu(g;\xi,\varepsilon) := \inf\{|z|_{\boldsymbol{\mathcal{Z}}} : z \in \mathcal{C}(g;\xi,\varepsilon)\} \tag{3.23}$$

for $g \in \boldsymbol{\mathcal{V}}$ and then, for $R > 0$, define

$$\beta(R) = \beta(R;\xi,\varepsilon) := \sup\{\nu(\mathbf{G}z;\xi,\varepsilon) : z \in \boldsymbol{\mathcal{Z}}_{ad},\ |z|_{\boldsymbol{\mathcal{Z}}} \le R\} \tag{3.24}$$

To have each $\nu(g)$ finite (for $\varepsilon > 0$ and $g \in \boldsymbol{\mathcal{G}}_*$) just means that $\xi \in \bar{\mathcal{K}}_g$ (for each $g \in \boldsymbol{\mathcal{G}}_*$) so (3.21) means that $\nu(g;\xi,\varepsilon) < \infty$ for each $\varepsilon > 0$ and every $\xi \in \mathcal{K}_0$. This does not yet mean that β will be finite and we will express the desired strengthening of (3.21) quantitatively in terms of $\nu(\cdot)$ and $\beta(\cdot)$.

Returning to the condition (3.20), we finish this section with some observations about the structure of the set $\boldsymbol{\mathcal{V}}_a = \boldsymbol{\mathcal{V}}_a(\boldsymbol{\mathcal{Z}}_{ad})$ given by

$$\boldsymbol{\mathcal{V}}_a := \{g \in \boldsymbol{\mathcal{V}} : \bar{\mathcal{K}}_g = \bar{\mathcal{K}}_0\} = \{g \in \boldsymbol{\mathcal{V}} : \mathbf{L}_{[\mathrm{T}]}g + \bar{\mathcal{K}}_0 = \bar{\mathcal{K}}_0\}, \tag{3.25}$$

noting that (3.20) just asserts that $\boldsymbol{\mathcal{G}}_* \subset \boldsymbol{\mathcal{V}}_a$. Since we always assume $\mathbf{L}_{[\mathrm{T}]}$ is continuous, we always have $\boldsymbol{\mathcal{V}}_a$ closed in $\boldsymbol{\mathcal{V}}$.

LEMMA 1. *$\boldsymbol{\mathcal{V}}_a$ is closed under addition and subtraction. If $\bar{\mathcal{K}}_0$ is convex, then $\boldsymbol{\mathcal{V}}_a$ is a* (*closed*) *subspace of $\boldsymbol{\mathcal{V}}$.*

Proof. Suppose $g, g' \in \boldsymbol{\mathcal{V}}_a$. We first wish to show that $\mathcal{K}_{\bar{g}} \subset \bar{\mathcal{K}}_0$ where $\bar{g} := g - g'$, i.e., for any $\bar{\xi} \in \mathcal{K}_{\bar{g}}$ and $\varepsilon > 0$ that there exists $\bar{z} \in \boldsymbol{\mathcal{Z}}_{ad}$ with $|\bar{\xi} - \mathbf{E}\bar{z}| \le \varepsilon$. To start, we have $\bar{\xi} = \mathbf{L}_{[\mathrm{T}]}\bar{g} + \mathbf{E}z_0 = \xi_1 - \mathbf{L}_{[\mathrm{T}]}g'$ with $\xi_1 := \mathbf{L}_{[\mathrm{T}]}g + \mathbf{E}z_0$. Since $\xi_1 \in \mathcal{K}_g$ and $g \in \boldsymbol{\mathcal{V}}_a$ gives $\mathcal{K}_g \subset \bar{\mathcal{K}}_0$, there must be $z_1 \in \boldsymbol{\mathcal{Z}}_{ad}$ such that $|\xi_1 - \mathbf{E}z_1| \le \varepsilon/2$. Now $\mathbf{E}z_1 \in \mathcal{K}_0$ and $g' \in \boldsymbol{\mathcal{V}}_a$ gives $\mathcal{K}_0 \subset \bar{\mathcal{K}}_{g'} = \mathbf{L}_{[\mathrm{T}]}g' + \bar{\mathcal{K}}_0$ so $(\mathbf{E}z_1 - \mathbf{L}_{[\mathrm{T}]}g') \in \bar{\mathcal{K}}_0$ and there must be $\bar{z} \in \boldsymbol{\mathcal{Z}}_{ad}$ with $|[\mathbf{E}z_1 - \mathbf{L}_{[\mathrm{T}]}g'] - \mathbf{E}\bar{z}| \le \varepsilon/2$. Since $\bar{\xi} - \mathbf{E}\bar{z} = (\xi_1 - \mathbf{E}z_1) + (\mathbf{E}z_1 - [\mathbf{L}_{[\mathrm{T}]}g' + \mathbf{E}\bar{z}])$, this gives $|\bar{\xi} - \mathbf{E}\bar{z}| \le \varepsilon$ as desired so $\bar{\xi} \in \bar{\mathcal{K}}_0$. This shows $\mathbf{L}_{[\mathrm{T}]}\bar{g} + \mathcal{K}_0 \subset \bar{\mathcal{K}}_0$ for $\bar{g} = g - g'$. Reversing the roles of g, g' gives $-\mathbf{L}_{[\mathrm{T}]}\bar{g} + \mathcal{K}_0 \subset \bar{\mathcal{K}}_0$ or $\mathcal{K}_0 \subset \bar{\mathcal{K}}_{\bar{g}}$. Combining gives $\bar{\mathcal{K}}_0 = \bar{\mathcal{K}}_{\bar{g}}$ so $\bar{g} \in \boldsymbol{\mathcal{V}}_a$ for $\bar{g} = g - g' \in \boldsymbol{\mathcal{V}}_a - \boldsymbol{\mathcal{V}}_a$, i.e., $\boldsymbol{\mathcal{V}}_a$ is closed under subtraction. Trivially, $0 \in \boldsymbol{\mathcal{V}}_a$ so $g' \in \boldsymbol{\mathcal{V}}_a$ gives $-g' \in \boldsymbol{\mathcal{V}}_a$ whence $\bar{g} = g - (-g') = g + g'$ is in $\boldsymbol{\mathcal{V}}_a$ for $g, g' \in \boldsymbol{\mathcal{V}}_a$, i.e., $\boldsymbol{\mathcal{V}}_a$ closed under addition also.

If we only show that convexity of $\bar{\mathcal{K}}_0$ implies that of $\boldsymbol{\mathcal{V}}_a$, then the algebraic closure above shows that $\boldsymbol{\mathcal{V}}_a$ is a subspace. Suppose, then, $\bar{g}$ is any convex combination of $\boldsymbol{\mathcal{V}}_a$ so $\bar{g} = \sum c_j g_j$ with $c_j > 0, \sum c_j = 1, g_j \in \boldsymbol{\mathcal{V}}_a$. For any $\xi \in \mathcal{K}_0$ we have $\mathbf{L}_{[\mathrm{T}]}\bar{g} + \xi = \sum c_j(\mathbf{L}_{[\mathrm{T}]}g_j + \xi)$. As each $g_j \in \boldsymbol{\mathcal{V}}_a$ we have each $(\mathbf{L}_{[\mathrm{T}]}g_j + \xi) \in \bar{\mathcal{K}}_0$ so convexity of $\bar{\mathcal{K}}_0$ gives $(\mathbf{L}_{[\mathrm{T}]}\bar{g} + \xi) \in \bar{\mathcal{K}}_0$. This, for each $\xi \in \mathcal{K}_0$, gives $\mathcal{K}_{\bar{g}} \subset \bar{\mathcal{K}}_0$. By Lemma 1 we have also $-\bar{g} = \sum c_j(-g_j)$ a convex combination of $\boldsymbol{\mathcal{V}}_a$ so $[\mathbf{L}_{[\mathrm{T}]}(-\bar{g}) + \xi] \in \bar{\mathcal{K}}_0$ for each $\xi \in \mathcal{K}_0$, i.e., $\xi \in [\mathbf{L}_{[\mathrm{T}]}\bar{g} + \bar{\mathcal{K}}_0] = \bar{\mathcal{K}}_{\bar{g}}$. Combining gives $\bar{\mathcal{K}}_0 = \bar{\mathcal{K}}_{\bar{g}}$ so $\bar{g} \in \boldsymbol{\mathcal{V}}_a$. □

Note that closure under addition shows that $\boldsymbol{\mathcal{V}}_a$ is always unbounded (except for the trivial case: $\boldsymbol{\mathcal{V}}_a = \{0\}$), so $\boldsymbol{\mathcal{Z}}_{ad}$ must then also be unbounded. Note that $\bar{\mathcal{K}}_0$ will certainly be convex if $\boldsymbol{\mathcal{Z}}_{ad}$ is convex.

4. Reachability. At this point we begin our analysis of approximate reachability by proving the trivial part of invariance: if, *for each affine perturbation* by $g \in \boldsymbol{\mathcal{G}}_* := \mathbf{G}\boldsymbol{\mathcal{Z}}_{ad}$ in (1.4), one can reach no state which is not already approximately reachable for (1.2), then the nonlinear perturbation (1.1) can also produce no new reachable states.

LEMMA 2. *Assume* $(\mathbf{H}_1)$(ii) *and the reachability inclusion*

$$\mathcal{K}_g \subset \bar{\mathcal{K}}_0 \textit{ for each } g \in \boldsymbol{\mathcal{G}}_*, \tag{4.26}$$

i.e., $\xi + \mathbf{L}_{[\mathrm{T}]}g \in \bar{\mathcal{K}}_0$ *for* $\xi \in \mathcal{K}_0$, $g \in \boldsymbol{\mathcal{G}}_*$. *Then* $\bar{\mathcal{K}}_\mathrm{F} \subset \bar{\mathcal{K}}_0$.

Proof. Since $\bar{\mathcal{K}}_\mathrm{F} = \bigcap_{\varepsilon>0}[\mathcal{K}_\mathrm{F} + \mathcal{B}_\varepsilon]$ with $\mathcal{B}_\varepsilon := \{\xi \in \mathcal{X} : |\xi| < \varepsilon\}$, we have $\xi \in \bar{\mathcal{K}}_\mathrm{F}$ if and only if for each $\varepsilon > 0$ there is some $\xi_\varepsilon = \mathbf{T}z_\varepsilon \in \mathcal{K}_\mathrm{F}$ with $|\xi_\varepsilon - \xi| < \varepsilon$. Setting $g_\varepsilon := \mathbf{G}z_\varepsilon$, we then have $\mathbf{T}z_\varepsilon = \mathbf{L}_{[\mathrm{T}]}g_\varepsilon + \mathbf{E}z_\varepsilon$ so $\xi_\varepsilon \in \mathcal{K}_{g_\varepsilon}$. This shows that we always have

$$\bar{\mathcal{K}}_\mathrm{F} \subset \bigcap_{\varepsilon>0} \bigcup_{g\in\boldsymbol{\mathcal{G}}_*} [\mathcal{K}_g + \mathcal{B}_\varepsilon] = \bigcap_{\varepsilon>0} \left[\left(\bigcup_{g\in\boldsymbol{\mathcal{G}}_*} \mathcal{K}_g \right) + \mathcal{B}_\varepsilon \right]. \tag{4.27}$$

Now, if one has (4.26), then the right side of (4.27) will be in $\bigcap_{\varepsilon>0}[\bar{\mathcal{K}}_0 + \mathcal{B}_\varepsilon] = \bar{\mathcal{K}}_0$ as asserted. □

For comparison we note that our previous work [11], [12], [15], [16] obtained essentially the following result.

THEOREM 1. *Let* $\boldsymbol{\mathcal{Z}}_{ad} = \boldsymbol{\mathcal{Z}}$ *be a Banach space* [3] *and assume* $(\mathbf{H}_1)$ *together with the growth condition*

$$|\mathbf{G}_\mathrm{F}z|_{\boldsymbol{\mathcal{V}}} \leq C_0 + C_1|z|^{\bar{r}}_{\boldsymbol{\mathcal{Z}}}. \tag{4.28}$$

for some $\bar{r} < 1$. *Assume the reachability invariance* (3.19): $\mathcal{K}_g = \mathcal{K}_0$ *for each* $g \in \boldsymbol{\mathcal{V}}$. *Then* $\mathcal{K}_\mathrm{F} = \mathcal{K}_0$, *i.e., the* exactly *reachable set is then invariant under the perturbation by* $\mathbf{F}$.

(*Of course this implies invariance of the approximately reachable set: we also have* $\bar{\mathcal{K}}_\mathrm{F} = \bar{\mathcal{K}}_0$.)

Proof. We only sketch the proof here to fix the ideas; for details, see [15]. From the inclusion $\mathcal{K}_g \subset \mathcal{K}_0$, we can argue as in Lemma 2 to have $\mathcal{K}_\mathrm{F} \subset \mathcal{K}_0$. The principal effort must go into showing that one also has the reverse inclusion $\mathcal{K}_0 \subset \mathcal{K}_\mathrm{F}$.

[3] In the previous papers $\boldsymbol{\mathcal{U}}_{ad} = \boldsymbol{\mathcal{U}}$ was a Banach space with $\mathbf{B}$ linear. The modification to the present formulation requires no essential change in the ideas at this point since we can topologize $\boldsymbol{\mathcal{Z}}$ by identification with $\boldsymbol{\mathcal{U}}/\mathcal{N}(\mathbf{B})$ when $\mathbf{B}$ is linear.

The proof begins by noting equivalence of (3.19) to the range condition

$$\mathbf{L}_{[\mathrm{T}]}\boldsymbol{\mathcal{V}} \subset \mathbf{E}\boldsymbol{\mathcal{Z}}. \tag{4.29}$$

Then the nullspace of the continuous linear map:

$$[g, \{z\}] \mapsto \mathbf{L}_{[\mathrm{T}]}g + \mathbf{E}z : \boldsymbol{\mathcal{V}} \times [\boldsymbol{\mathcal{Z}}/\mathcal{N}(\mathbf{E})] \to \mathcal{X}$$

will be the graph of a linear operator: $g \mapsto \{z\}$ which is bounded by the Closed Graph Theorem. By the Michael Selection Theorem [10], there is a continuous right inverse to the canonical projection: $z \mapsto \{z\} : \boldsymbol{\mathcal{Z}} \to [\boldsymbol{\mathcal{Z}}/\mathcal{N}(\mathbf{E})]$ which may be taken to be of linear growth.

Composing, for any fixed $\xi \in \mathcal{K}_0$, there is a continuous selection $\mathbf{C} = \mathbf{C}_\xi : \boldsymbol{\mathcal{V}} \to \boldsymbol{\mathcal{Z}}$ satisfying

$$\mathbf{L}_{[\mathrm{T}]}g + \mathbf{E}[\mathbf{C}g] = \xi \tag{4.30}$$

and a linear growth condition: $|\mathbf{C}g|_{\boldsymbol{\mathcal{Z}}} = \mathcal{O}(|g|_{\boldsymbol{\mathcal{V}}})$. It is an important point that $\mathbf{C}$ depends only on the range inclusion (4.29) and not at all on $\mathbf{F}$.

We then consider the map: $\boldsymbol{\mathcal{Z}} \to \boldsymbol{\mathcal{Z}}$ given by

$$\mathbf{C}\mathbf{G}_{\mathrm{F}} : z \mapsto g := \mathbf{G}_{\mathrm{F}}z \mapsto \hat{z} := \mathbf{C}g \tag{4.31}$$

and note that this can be restricted to an invariant ball by using the linear growth of $\mathbf{C}$ and the sublinear growth condition in (4.28). Using $(\mathbf{H}_1)$(iii), the Schauder Fixpoint Theorem applies to give the desired result since using a fixpoint $\hat{z}$ of $\mathbf{C}_\xi\mathbf{G}$ in (3.15) just gives $\mathbf{T}_{\mathrm{F}}\hat{z} = \xi$. □

Our intention here is to use, for the analysis of the approximately reachable set, a modified version of the argument used above for the exactly reachable set. We already have Lemma 2 and have noted that the real problem is to determine when $\bar{\mathcal{K}}_0 \subset \bar{\mathcal{K}}_{\mathrm{F}}$, which is just the question (**Q**) of the Introduction. It is not difficult to see, much as for (4.27), that we have

$$\bar{\mathcal{K}}_{\mathrm{F}} = \bigcap_{\varepsilon>0} \{\xi \in \mathcal{X} : \Phi(\xi,\varepsilon) \neq \emptyset\} \tag{4.32}$$

where $\Phi(\xi,\varepsilon) := \{z \in \boldsymbol{\mathcal{Z}}_{ad} : z \in \mathcal{C}(\mathbf{G}z;\xi,\varepsilon)\}$. Our argument, then, is to show that the set-valued map: $z \mapsto \mathcal{C}(\mathbf{G}z;\xi,\varepsilon)$ has a fixpoint (so $\Phi(\xi,\varepsilon)$ is nonempty) for $\xi \in \mathcal{K}_0$ and all $\varepsilon > 0$. We take the inclusion $\mathcal{K}_0 \subset \bar{\mathcal{K}}_g$ of (3.20) as a starting point but, again as noted, we will be forced to strengthen this.

The critical difficulty with this program is that (4.29), with $\boldsymbol{\mathcal{Z}}_{ad}$ a Banach space, was essential for the applicability of the theorems used above to obtain the essential continuity and linear growth for $\mathbf{C}_\xi$ and so, with the growth condition (4.28), existence of an invariant ball for the mapping $\mathbf{C}_\xi\mathbf{G}_{\mathrm{F}}$. The various results in this paper correspond to ways of handling this difficulty.

We have already made our *third basic observation*: for present purposes we may not only fix ξ (arbitrary $\xi \in \mathcal{K}_0$) but also ε (arbitrary $\varepsilon > 0$): now considering $\mathbf{C} = \mathbf{C}_{\xi,\varepsilon} : \boldsymbol{\mathcal{G}} \to \boldsymbol{\mathcal{Z}}_{ad}$, we weaken (4.30) to require only that

$$\mathbf{C}_{\xi,\varepsilon}(g) \in \mathcal{C}(g;\xi,\varepsilon) \text{ so } |\xi - (\mathbf{L}_{[\mathrm{T}]}g + \mathbf{E}[\mathbf{C}g])|_{\mathcal{X}} \leq \varepsilon. \tag{4.33}$$

One easily sees that having $\mathcal{C}(g;\xi,\varepsilon)$ nonempty (for each $\xi \in \mathcal{K}_0$ and each $\varepsilon > 0$) is precisely equivalent to the inclusion $\mathcal{K}_0 \subset \bar{\mathcal{K}}_g$. Given this and the compactness of $\mathbf{G}_{\mathrm{F}}$, it is not too difficult to construct $\mathbf{C} = \mathbf{C}_{\xi,\varepsilon}$ continuous giving (4.33) for each $g \in \boldsymbol{\mathcal{G}}_*$. Unfortunately, without strengthening this condition we cannot obtain a growth rate for $\mathbf{C} = \mathbf{C}_{\xi,\varepsilon}$ which, with (4.28), would give a bounded invariant set. (Under the strong assumption that $\mathbf{E}\boldsymbol{\mathcal{G}}_*$ is known to be precompact in $\mathcal{X}$, we do, however, have Theorem 6, below.) Our basic approach is embodied in the following theorems.

THEOREM 2. *Assume* $(\mathbf{H}_1)$. *Let* $\boldsymbol{\mathcal{Z}}_{ad}$ *be convex and, for some* $\xi \in \mathcal{X}$ *and* $\varepsilon > 0$, *assume that there is some* $R = R(\xi, \varepsilon)$ *such that*

$$\beta(R; \xi, \varepsilon) < R. \tag{4.34}$$

Then there is some $\bar{z} \in \boldsymbol{\mathcal{Z}}_{ad}$ *(with* $|\bar{z}| \leq R$*) such that* $|\xi - \mathbf{T}_{\mathrm{F}}\bar{z}| \leq 2\varepsilon$.

Proof. By $(\mathbf{H}_1)$ we have $\mathbf{G}$, $\mathbf{L}$, $\mathbf{T}_{\mathrm{F}}$ continuous so there is no loss of generality in taking $\boldsymbol{\mathcal{Z}}_{ad}$ closed in $\boldsymbol{\mathcal{Z}}$ (i.e., replacing it by its closure). Now define

$$\boldsymbol{\mathcal{Z}}_R := \{z \in \boldsymbol{\mathcal{Z}}_{ad} : |z| \leq R\}, \quad \boldsymbol{\mathcal{G}}_R := \mathbf{G}\boldsymbol{\mathcal{Z}}_R = \{\mathbf{G}z : z \in \boldsymbol{\mathcal{Z}}_R\}. \tag{4.35}$$

Note that (i) $\boldsymbol{\mathcal{Z}}_R$ is closed and convex in $\boldsymbol{\mathcal{Z}}$, (ii) $\boldsymbol{\mathcal{G}}_R$ is precompact in $\boldsymbol{\mathcal{V}}$ by $(\mathbf{H}_1)$(iii), and (iii) $\mathcal{C}_R(g) := \boldsymbol{\mathcal{Z}}_R \cap \mathcal{C}(g; \xi, \varepsilon)$ is nonempty for each $g \in \boldsymbol{\mathcal{G}}_R$ since $\nu(g) \leq \beta(R) < R$. By (ii), we can find a finite set of "centers" $\{g_j : j = 1, \cdots, J\}$ such that

$$\min_j\{|g - g_j|\} \leq \delta := \varepsilon/2\|\mathbf{L}_{[\mathrm{T}]}\| \text{ for each } g \in \boldsymbol{\mathcal{G}}_R \tag{4.36}$$

and by (iii), we can find $z_j \in \mathcal{C}_R(g_j)$ for each j. A standard construction gives a continuous partition of unity subordinate to the covering of $\boldsymbol{\mathcal{G}}_R$ by 2δ-balls centered at $\{g_j\}$, i.e., continuous scalar functions φ_j on $\boldsymbol{\mathcal{V}}$ such that

$$\varphi_j \geq 0, \ \sum \varphi_j(g) = 1 \text{ for } g \in \boldsymbol{\mathcal{G}}_R, \quad \varphi_j(g) > 0 \Longrightarrow |g - g_j| \leq 2\delta.$$

We now define $\mathbf{C} = \mathbf{C}_{\xi,\varepsilon}$ by

$$\mathbf{C}g := \sum \varphi_j(g) z_j$$

and note that this gives $\mathbf{C} : \boldsymbol{\mathcal{G}}_R \to \boldsymbol{\mathcal{Z}}_R$ since $\boldsymbol{\mathcal{Z}}_R$ is convex. Clearly $\mathbf{C}$ is continuous and, as each $z_j \in \mathcal{C}(g_j)$, a simple computation from (3.22) gives

$$|\xi - [\mathbf{L}_{[\mathrm{T}]}g + \mathbf{E}\mathbf{C}g]|_{\mathcal{X}} \leq 2\varepsilon \quad \text{for each } g \in \boldsymbol{\mathcal{G}}_R. \tag{4.37}$$

From $(\mathbf{H}_1)$ we have $\mathbf{C}\mathbf{G} : \boldsymbol{\mathcal{Z}}_R \to \boldsymbol{\mathcal{Z}}_R$ continuous and compact so, applying the Schauder Fixpoint Theorem, there is a fixpoint $\bar{z} \in \boldsymbol{\mathcal{Z}}_R$, i.e., we have $\mathbf{C}\bar{g} = \bar{z}$ for $\bar{g} = \mathbf{G}\bar{z}$. Putting $g = \bar{g}$ gives $\mathbf{L}_{[\mathrm{T}]}g + \mathbf{E}\mathbf{C}g = \mathbf{L}_{[\mathrm{T}]}\mathbf{G}\bar{z} + \mathbf{E}\bar{z} =: \mathbf{T}_{\mathrm{F}}\bar{z}$ so (4.37) gives $|\xi - \mathbf{T}_{\mathrm{F}}\bar{z}| \leq 2\varepsilon$ as desired. □

THEOREM 3. *Assume* $(\mathbf{H}_1)$ *with* (4.28) *and that* $\boldsymbol{\mathcal{Z}}_{ad}$ *is convex. Now suppose, for some* $\xi \in \mathcal{K}_0$ *and each* $\varepsilon > 0$, *that one has a growth rate*

$$\nu(g) \leq \tilde{C}_0 + \tilde{C}_1|g|_{\boldsymbol{\mathcal{V}}}^{\tilde{r}} \quad \textit{for } g \in \boldsymbol{\mathcal{G}}_* \tag{4.38}$$

with $\tilde{r} < 1/\bar{r}$. *(Here,* ν *is defined as in* (3.23)*; the numbers* $\tilde{C}_0, \tilde{C}_1, \tilde{r}$ *may depend on* ξ, ε.*) Then one has* $\xi \in \bar{\mathcal{K}}_{\mathrm{F}}$.

Proof. Substituting (4.28) in (4.38) gives

$$\beta(R) \leq \tilde{C}_0 + \tilde{C}_1[C_0 + C_1 R^{\bar{r}}]^{\tilde{r}} = \mathcal{O}(R^{\bar{r}\tilde{r}}) = o(R)$$

as $\bar{r}\tilde{r} < 1$. Hence one can always find $R = R(\xi, \varepsilon)$ for which $\beta(R) < R$ so Theorem 2 applies. □

Note that if we have this for *every* $\xi \in \mathcal{K}_0$ and also have the hypotheses of Lemma 2, then we have shown, as in the title, *invariance of the approximately reachable set* under the nonlinear perturbation $\mathbf{F}$, i.e., $\bar{\mathcal{K}}_{\mathrm{F}} = \bar{\mathcal{K}}_0$.

The condition (4.28) is the same growth condition for the nonlinearity as was imposed in Theorem 1. Note also that the condition (4.38) implies, in particular, that ν is finite and so quantitatively strengthens the simple requirement of approximate reachability for each (1.4):

$$\xi \subset \bar{\mathcal{K}}_g = \mathbf{L}_{[\mathrm{T}]}g + \bar{\mathcal{K}}_0 \text{ for each } g \in \boldsymbol{\mathcal{G}}_* := \{\mathbf{G}_{\mathrm{F}}z : z \in \boldsymbol{\mathcal{Z}}_{ad}\} \tag{4.39}$$

which, for every $\xi \in \mathcal{K}_0$, would just be (3.21). We can also use this approach — more-or-less unchanged, with proper formulation — in a somewhat more complicated nonlinear setting in which the linear operator $\mathbf{A}$ is also taken to depend on x, u. Without developing this further at the present time, we provide the formulation here in the hope that it will also add to the reader's insight into the essential aspects of the original argument above. Clearly one might continue to extend the same basic ideas.

We now suppose that we wish to analyze approximate reachability for an equation of the form

$$\dot{x} = \mathbf{A}(x, u)x + f(x, u) + \mathbf{B}u, \qquad u \in \boldsymbol{\mathcal{U}}_{ad} \tag{4.40}$$

by comparison with similar considerations for the family of (presumably simpler) equations (1.4); now, $\mathbf{A}$ ranges over a suitable set $\boldsymbol{\mathcal{A}}$ of linear operators and g ranges over $\boldsymbol{\mathcal{G}}$. As above, we assume that $\mathbf{A}(x_u, u) \in \boldsymbol{\mathcal{A}}$ and $f(x_u, u) \in \boldsymbol{\mathcal{G}}$ for each solution x_u of (4.40) as u ranges over $\boldsymbol{\mathcal{U}}_{ad}$ and that that solutions are here also taken in $\boldsymbol{\mathcal{X}} := C([0,T] \to \mathcal{X})$ where $\mathcal{X}$ is the state space. We assume that there are well-defined maps

$$\begin{array}{rccl} \mathbf{M}: & \boldsymbol{\mathcal{U}}_{ad} \to \boldsymbol{\mathcal{A}} & : & u \mapsto \mathbf{A}(x_u, u), \\ \mathbf{L}_{[\mathrm{T}]}: & \boldsymbol{\mathcal{U}}_{ad} \times \boldsymbol{\mathcal{A}} \times \boldsymbol{\mathcal{G}} \to \mathcal{X} & : & [u, \mathbf{A}, g] \mapsto x_u(T) \end{array}$$

given by (4.40) and (1.4), respectively. We will take $\boldsymbol{\mathcal{U}}_{ad}$ to be a convex subset of a suitable Banach space and take $\boldsymbol{\mathcal{A}}$, $\boldsymbol{\mathcal{G}}$ to be compact Hausdorff spaces; we assume the topologies can be taken so that $\mathbf{L}_{[\mathrm{T}]}$ and $\mathbf{M}$ are continuous. Much as before, we then introduce

$$\begin{aligned} \mathcal{C}(\mathbf{A}, g; \xi, \varepsilon) &= \{u \in \boldsymbol{\mathcal{U}}_{ad} : |\xi - \mathbf{L}_{[\mathrm{T}]}[u, \mathbf{A}, g]| < \varepsilon\}, \\ \nu(\mathbf{A}, g; \xi, \varepsilon) &= \inf\{|u| : u \in \mathcal{C}(\mathbf{A}, g; \xi, \varepsilon)\}, \\ R(\xi, \varepsilon) &= \sup\{\nu(\mathbf{A}, g; \xi, \varepsilon) : \mathbf{A} \in \boldsymbol{\mathcal{A}},\ g \in \boldsymbol{\mathcal{G}}\}. \end{aligned}$$

For $\nu(\mathbf{A}, g; \xi, \varepsilon)$ to be finite just means that $\mathcal{C}(\mathbf{A}, g; \xi, \varepsilon)$ is nonempty and this will be the case, for the given $\xi \in \mathcal{X}$ and every $\varepsilon > 0$, if and only if this ξ is approximately reachable for (1.4) using this $\mathbf{A} \in \boldsymbol{\mathcal{A}}$. This, of course, would not be sufficient in itself to ensure finiteness of $R(\xi, \varepsilon)$ unless, for example, one were to know that $\nu(\cdot; \xi, \varepsilon)$ would be upper semicontinuous on the (assumed) compact set $\boldsymbol{\mathcal{A}} \times \boldsymbol{\mathcal{G}}$.

THEOREM 4. *Let the setting be as assumed above and suppose, for some fixed $\xi \in \mathcal{X}$, that we have $R(\xi, \varepsilon) < \infty$ for each $\varepsilon > 0$. Then this ξ is approximately reachable for* (4.40).

Proof. Choose some $R > R(\xi, \varepsilon)$ and restrict attention to the ball $\boldsymbol{\mathcal{U}}_R := \{u \in \boldsymbol{\mathcal{U}}_{ad} : |u| \le R\}$. We can then construct a continuous map $\mathbf{C} = \mathbf{C}_{\xi,\varepsilon} : [\boldsymbol{\mathcal{A}} \times \boldsymbol{\mathcal{G}}] \to \boldsymbol{\mathcal{U}}_R$ such that, corresponding to (4.33), we have

$$|\xi - \mathbf{L}_{[\mathrm{T}]}[\mathbf{C}(\mathbf{A}, g), \mathbf{A}, g]|_{\mathcal{X}} \le 2\varepsilon$$

for each $[\mathbf{A}, g] \in \boldsymbol{\mathcal{A}} \times \boldsymbol{\mathcal{G}}$. We omit detail since this is much as in the proof of Theorem 2 — what corresponds to (4.36) is that we can find, for each $[\mathbf{A}, g]$, a control $\tilde{u} \in \boldsymbol{\mathcal{U}}_R \cap \mathcal{C}(\mathbf{A}, g; \xi, \varepsilon)$ and then a neighborhood $\mathcal{N} = \mathcal{N}(\mathbf{A}, g)$ on which the continuous function $\mathbf{L}_{[\mathrm{T}]}[\tilde{u}, \cdot]$ varies by no more than ε in $\mathcal{X}$-norm; there is then a finite subcovering by such neighborhoods, and one can find a corresponding partition of unity.

Note that the continuity of $\mathbf{G}$ and the compactness of $\boldsymbol{\mathcal{A}} \times \boldsymbol{\mathcal{G}}$ make the composition $\mathbf{CG}$ a continuous, compact selfmap of the bounded convex set $\boldsymbol{\mathcal{U}}_R$. The Schauder Fixpoint Theorem applies to ensure a fixpoint $\hat{u}$ for which (4.40) gives $|\xi - x_{\hat{u}}(T)| \le 2\varepsilon$. This, for each $\varepsilon > 0$, ensures the approximate reachability of this ξ for (4.40). □

5. Further results. Having verified $(\mathbf{H}_1)$, the principal problem in applying the general results above is, of course, the difficulty in verifying a condition such as (4.38) to enable one to restrict attention to some $\boldsymbol{\mathcal{Z}}_R$. There are, however, certain cases in which one can proceed. We note one in which $\boldsymbol{\mathcal{Z}}_{ad}$ is the whole Banach space $\boldsymbol{\mathcal{Z}}$ and another in which the nonlinearity $\mathbf{F}$ gives only a compactly restricted perturbation of the final value.

Since we only consider $\nu(g; \xi_0, \varepsilon)$ for $\xi_0 \in \mathcal{K}_0$ so $\xi_0 = \mathbf{T}z_0$, we can introduce

$$\bar{\nu}(\bar{\xi}; \varepsilon) := \inf\{|z'| : z' \in [z_0 - \boldsymbol{\mathcal{Z}}_{ad}],\ |\bar{\xi} - \mathbf{T}z'| \le \varepsilon\}$$

and have $\nu(g; \mathbf{T}z_0, \varepsilon) = \bar{\nu}(\mathbf{L}_{[\mathrm{T}]}g, \varepsilon)$. Observe that if we consider $\boldsymbol{\mathcal{Z}}_{ad} = \boldsymbol{\mathcal{Z}}$, then scaling gives $\bar{\nu}(\lambda\bar{\xi}, \varepsilon) = \lambda\bar{\nu}(\bar{\xi}, \varepsilon/\lambda)$ so (4.38) is equivalent to requiring that

$$\nu_*^{\vartheta}(\bar{\xi}) := \limsup_{\varepsilon \to 0} \{\inf_{z} \{\varepsilon^{-(1-\vartheta)}|z| : z \in \boldsymbol{\mathcal{Z}} \text{ and } |\bar{\xi} - \mathbf{E}z|_{\mathcal{X}} \le \varepsilon\}\}$$

should be bounded for $\bar{\xi} \in \{\mathbf{L}_{[\mathrm{T}]}g : g$ in some $\boldsymbol{\mathcal{V}}$-bounded subset of $\boldsymbol{\mathcal{G}}_*\}$. It is possible to show that ν_*^{ϑ} is actually a norm intermediate between the $\mathcal{X}$-norm on $\bar{\mathcal{K}}_0$ ($\vartheta = 0$) and the obvious induced norm: $|\xi|_1 := \inf\{|z| : \mathbf{T}z = \xi\}$ on $\mathcal{K}_0$ ($\vartheta = 1$). Thus, finiteness of $\nu_*^{\vartheta}(\mathbf{L}_{[\mathrm{T}]}g)$ is a stronger condition than just requiring $\mathbf{L}_{[\mathrm{T}]}g \in \bar{\mathcal{K}}_0$ but is weaker than the exact reachability condition of Theorem 2, i.e., that $\mathbf{L}_{[\mathrm{T}]}g \in \mathcal{K}_0$. We will not analyze ν_*^{ϑ} directly but will, instead, use the established theory Banach space interpolation (cf., e.g., [4]).

THEOREM 5. *Suppose* $\boldsymbol{\mathcal{Z}}_{ad} = \boldsymbol{\mathcal{Z}}$ *so* $\bar{\mathcal{K}}_0 =: \mathcal{X}_0$ *is a* (*closed*) *subspace of* $\mathcal{X}$. *Assume* $(\mathbf{H}_1)$ *and suppose*[4] *that, for some* $\vartheta > 0$, *one has*

$$\mathbf{L}_{[\mathrm{T}]}g \in \mathcal{X}_{\vartheta} \quad \textit{for each } g \in \boldsymbol{\mathcal{V}} \tag{5.41}$$

where $\mathcal{X}_{\vartheta}$ *is an interpolation space* $[\mathcal{X}_0, \mathcal{X}_1]_{\vartheta}$ *with* $\mathcal{X}_1 := \mathcal{K}_0$ (*taken with the norm:* $|\xi|_1 := \inf\{|z| : \mathbf{E}z = \xi$ *for* $\xi \in \mathcal{X}_1 = \mathcal{K}_0\}$*; note that here* $|z|$ *is the* $\boldsymbol{\mathcal{Z}}$-*norm*) *and* $\mathcal{X}_0 := \bar{\mathcal{K}}_0$ *with the* $\mathcal{X}$-*norm. Let* $\mathbf{F}$ *be a nonlinearity satisfying the growth condition* (4.28) *for some* $\bar{r} < \vartheta$. *Then the approximately reachable set* $\bar{\mathcal{K}}_{\mathrm{F}}$ *is precisely* $\bar{\mathcal{K}}_0$.

[4] Given $\mathbf{F}$, the hypothesis (5.41) with $\vartheta > \bar{r}$ is somewhere between taking $\vartheta = 0$, which just reduces to the (inadequate) hypothesis (4.39), and taking $\vartheta = 1$ which is equivalent to the exact reachability inclusion (4.29) used in Theorem 1. Note that it is easiest to obtain (5.41) if one takes $\boldsymbol{\mathcal{V}}$ as small as possible consistent with $(\mathbf{H}_1)$.

Proof. Since (5.41) gives $\mathbf{L}_{[\mathrm{T}]}\boldsymbol{\mathcal{V}} \subset \mathcal{X}_1 \subset \mathcal{X}_0$ and these are subspaces, we have (4.26) so $\bar{\mathcal{K}}_{\mathrm{F}} \subset \bar{\mathcal{K}}_0$ by Lemma 2. As usual, we are primarily concerned with the complementary inclusion.

While there are various possible interpolation functors, the extremal property of the K-functor (see, e.g., Theorem 3.9.1 of [4]) gives a uniform estimate:

$$s^{-\vartheta}K(s;\xi) \leq C|\xi|_\vartheta \quad (s > 0,\ \xi \in \mathcal{X}_\vartheta) \tag{5.42}$$

(C depending on the choice of $|\cdot|_\vartheta$) where the function $K(\cdot)$ is defined by

$$\begin{aligned} K(s;\xi) &:= \inf\{|\xi_0|_{\mathcal{X}} + s|\xi_1|_1 : \ \xi_0 + \xi_1 = \xi,\ \xi_1 \in \mathcal{X}_1\} \\ &= \inf\{s|z| + |\xi - \mathbf{E}z| : \ z \in \boldsymbol{\mathcal{Z}}\}. \end{aligned} \tag{5.43}$$

Fixing $\varepsilon > 0$, define

$$\omega(\nu) = \omega(\nu;\varepsilon) := [C\,\varepsilon^{-(1-\vartheta)}]^{1/\vartheta}\ \nu^{1/\vartheta}.$$

For any $\xi \in \mathcal{X}_\vartheta$ set $\nu := |\xi|_\vartheta$ and consider $s = \varepsilon/\omega$ in (5.42), (5.43) with $\omega > \omega(\nu)$.

From (5.42) this gives $K(s;\xi) < \varepsilon$ so, from (5.43), there exists $z \in \boldsymbol{\mathcal{Z}}$ such that

$$|\xi - \mathbf{E}z| < \varepsilon, \qquad |z| < \omega.$$

(Here the first is the $\mathcal{X}$-norm while the second is the $\boldsymbol{\mathcal{Z}}$-norm.) Since we may take ω arbitrarily close to $\omega(\nu)$, this shows:

$$\inf\{|z| : |\xi - \mathbf{E}z| \leq \varepsilon\} \leq [C\varepsilon^{-(1-\vartheta)}]^{1/\vartheta}|\xi|_\vartheta^{1/\vartheta} \tag{5.44}$$

for $\xi \in \mathcal{X}_\vartheta$.

Note that (5.41) implies, by the Closed Graph Theorem, continuity of $\mathbf{L}_{[\mathrm{T}]}$ as a linear operator from $\boldsymbol{\mathcal{V}}$ to $\mathcal{X}_\vartheta$, i.e., existence of a constant $\bar{C}$ such that $|\mathbf{L}_{[\mathrm{T}]}g|_\vartheta \leq \bar{C}|g|$ ($|g|$ is the $\boldsymbol{\mathcal{V}}$-norm). Now fix $\xi_0 = \mathbf{E}z_0 \in \mathcal{K}_0 = \mathcal{X}_1$ and, letting $\xi = \mathbf{L}_{[\mathrm{T}]}g$ in (5.44), note that $|\xi - \mathbf{E}z| \leq \varepsilon$ if and only if

$$|\xi_0 - [\mathbf{L}_{[\mathrm{T}]}g + \mathbf{E}(z_0 - z)]|_{\mathcal{X}} \leq \varepsilon \ \text{ so } z_0 - z =: z' \in \mathcal{C}(g;\xi_0,\varepsilon).$$

From (5.44), any $\omega > \omega(|\mathbf{L}_{[\mathrm{T}]}g|_\vartheta)$ can be used to estimate z' so

$$\begin{aligned} \nu(g;\xi_0,\varepsilon) &= \inf\{|z'| : z' = z_0 - z \in \mathcal{C}(g;\xi_0,\varepsilon)\} \\ &\leq |z_0| + \inf\{|z| : z_0 - z \in \mathcal{C}(g;\xi_0,\varepsilon)\} \\ &\leq |z_0| + [C\varepsilon^{-(1-\vartheta)}]^{1/\vartheta}[\bar{C}|g|]^{1/\vartheta} \end{aligned} \tag{5.45}$$

We recognize this as (4.38) with $\tilde{r} = 1/\vartheta$; the assumption $\vartheta > \bar{r}$ gives $\tilde{r} < 1/\bar{r}$. Thus Theorem 3 applies to show $\mathcal{K}_0 \subset \bar{\mathcal{K}}_{\mathrm{F}}$ and one has the desired invariance: $\bar{\mathcal{K}}_{\mathrm{F}} = \bar{\mathcal{K}}_0$. □

COROLLARY. *Suppose* $\boldsymbol{\mathcal{V}}_0$ *is any space for which* $\mathbf{L}_{[\mathrm{T}]} : \boldsymbol{\mathcal{V}}_0 \to \mathcal{X}$ *is continuous and* $\boldsymbol{\mathcal{V}}_1$ *is any space for which the exact reachability condition holds:* $\{\mathbf{L}_{[\mathrm{T}]}g : g \in \boldsymbol{\mathcal{V}}_1\} \subset \mathcal{K}_0$. *Assume* ($\mathbf{H}_1$) *with* $\boldsymbol{\mathcal{V}}$ *taken as* $\boldsymbol{\mathcal{V}}_\vartheta := [\boldsymbol{\mathcal{V}}_0, \boldsymbol{\mathcal{V}}_1]_\vartheta$ *for some* $\vartheta > \bar{r}$; *assume* (4.26) *and* (4.28). *Then one has* $\bar{\mathcal{K}}_{\mathrm{F}} = \bar{\mathcal{K}}_0$.

Proof. Let $\mathbf{L}^0_{[\mathrm{T}]}$ be $\mathbf{L}_{[\mathrm{T}]} : \boldsymbol{\mathcal{V}}_0 \to \mathcal{X}_0 := \bar{\mathcal{K}}_0$ and let $\mathbf{L}^1_{[\mathrm{T}]}$ be $\mathbf{L}_{[\mathrm{T}]} : \boldsymbol{\mathcal{V}}_1 \to \mathcal{X}_1 := \mathcal{K}_0$; the latter is bounded by the Closed Graph Theorem since $\mathbf{L}_{[\mathrm{T}]}\boldsymbol{\mathcal{V}}_1 \subset \mathcal{X}_1$. Then boundedness of $\mathbf{T}^\vartheta : \boldsymbol{\mathcal{V}}_\vartheta \to \mathcal{X}_\vartheta := [\mathcal{X}_0, \mathcal{X}_1]_\vartheta$ follows from interpolation theory [4] and Theorem 5 applies. □

A somewhat modified fixpoint argument provides our final result.

THEOREM 6. *Assume* $(\mathbf{H}_1)$(i),(ii) *and* (4.26). *Let* $\boldsymbol{\mathcal{Z}}_{ad}$ *be convex and closed in* $\boldsymbol{\mathcal{Z}}$ *and assume we have* (3.21) *for each* $g \in \boldsymbol{\mathcal{G}}_* := \{\mathbf{G}_F z : z \in \boldsymbol{\mathcal{Z}}_{ad}\}$. *Finally, we assume that* $\mathbf{L}_{[T]}\boldsymbol{\mathcal{G}}_* := \{\mathbf{L}_{[T]}\mathbf{G}_F z : z \in \boldsymbol{\mathcal{Z}}_{ad}\}$ *is precompact in* $\mathcal{X}$. *Then, one has* $\bar{\mathcal{K}}_F = \bar{\mathcal{K}}_0$.

Proof. By Lemma 2 we have $\bar{\mathcal{K}}_F \subset \bar{\mathcal{K}}_0$ and, as earlier, need only show $\xi \in \bar{\mathcal{K}}_F$ for each fixed $\xi = \xi_0 \in \bar{\mathcal{K}}_0$.

Note that (3.21) just means that $[\xi_0 - \mathbf{L}_{[T]}g] \in \bar{\mathcal{K}}_0$ and this, for each $g \in \mathbf{G}_*$, means $[\xi_0 - \mathbf{L}_{[T]}\mathbf{G}_*] \subset \bar{\mathcal{K}}_0$. As $\bar{\mathcal{K}}_0 := \overline{\mathbf{E}\boldsymbol{\mathcal{Z}}_{ad}}$ is convex, since $\boldsymbol{\mathcal{Z}}_{ad}$ is and $\mathbf{E}$ is linear, we have $\mathcal{X}_* := \overline{co}(\xi_0 - \mathbf{L}_{[T]}\boldsymbol{\mathcal{G}}_*) \subset \bar{\mathcal{K}}_0$. On the other hand, we note that $\mathcal{X}_0 := \overline{co}(\mathbf{L}_{[T]}\boldsymbol{\mathcal{G}}_*)$ is convex and is compact by the precompactness assumption so $\mathcal{X}_*$ is compact since we easily see $\mathcal{X}_* := \xi_0 - \mathcal{X}_0$.

Given any $\varepsilon > 0$, one can then find a finite covering of $\mathcal{X}_*$ by ε-balls which may be taken centered at $\{\xi_j : j = 1, \cdots, J\}$ with each $\xi_j \in \mathcal{K}_0 \cap \mathcal{X}_*$ so there exist $z_j \in \boldsymbol{\mathcal{Z}}_{ad}$ such that $\mathbf{E}z_j = \xi_j$. As in the proof of Theorem 2,we can find a continuous partition of unity subordinate to this covering:

$$\varphi_j \geq 0, \quad \sum \varphi_j \equiv 1 \text{ on } \mathcal{X}_*, \quad \varphi_j(\xi) \neq 0 \Longrightarrow |\xi - \xi_j| < \varepsilon.$$

and then define $\mathbf{C} = \mathbf{C}_\varepsilon$ by

$$\mathbf{C}\xi := \sum \varphi_j(\xi) z_j,$$

noting that $\mathbf{C}\xi \in \boldsymbol{\mathcal{Z}}_{ad}$ for $\xi \in \mathcal{X}_*$ by the assumed convexity of $\boldsymbol{\mathcal{Z}}_{ad}$. Clearly $\mathbf{C} : \mathcal{X}_* \to \boldsymbol{\mathcal{Z}}_{ad}$ is continuous and, as earlier, a simple computation shows that

$$|\xi - \mathbf{EC}\xi| \leq \varepsilon \quad \text{for } \xi \in \mathcal{X}_*. \tag{5.46}$$

As already noted, for any $z \in \boldsymbol{\mathcal{Z}}_{ad}$ we have $[\xi - \mathbf{L}_{[T]}\mathbf{G}z] \in \mathcal{X}_*$ so the map:

$$\xi \mapsto z := \mathbf{C}\xi \mapsto [\xi - \mathbf{L}_{[T]}\mathbf{G}z] \tag{5.47}$$

is a continuous selfmap of the compact, convex set $\mathcal{X}_*$.

By the Schauder Fixpoint Theorem this map has a fixpoint $\bar{\xi}$ so, setting $\bar{z} := \mathbf{C}\bar{\xi} \in \boldsymbol{\mathcal{Z}}_{ad}$ we have $\bar{\xi} = \xi - \mathbf{L}_{[T]}\mathbf{G}\bar{z}$. Using (5.46), we have $|\xi - \mathbf{T}_F\bar{z}|_{\mathcal{X}} = |\xi - [\mathbf{L}_{[T]}\mathbf{G}\bar{z} + \mathbf{T}\bar{z}]|_{\mathcal{X}} = |\bar{\xi} - \mathbf{E}\bar{z}|_{\mathcal{X}} \leq \varepsilon$. Since this is possible for each $\varepsilon > 0$, we have $\xi \in \bar{\mathcal{K}}_F \subset \bar{\mathcal{K}}_F$. Since that holds for each $\xi \in \bar{\mathcal{K}}_0$, we have $\bar{\mathcal{K}}_0 \subset \bar{\mathcal{K}}_F$. □

6. Approaches to the hypotheses.

Finally, we consider the verification of the abstract hypothesis $(\mathbf{H}_1)$ for more concrete settings — e.g., such as those presented in §2. We begin by noting that in connection with $(\mathbf{H}_1)$ (ii), (iii) there may be an advantage in introducing an auxiliary function space $\boldsymbol{\mathcal{Y}}$.

LEMMA 3. *Suppose* $\boldsymbol{\mathcal{Z}}$ *embeds continuously in a space* $\boldsymbol{\mathcal{Y}}$ *which is compatible with* $\boldsymbol{\mathcal{X}}$ *in the sense that the set* $\boldsymbol{\mathcal{Y}} \cap \boldsymbol{\mathcal{X}}$ *is dense both in* $\boldsymbol{\mathcal{Y}}$ *and in* $\boldsymbol{\mathcal{X}}$ *with*

$$y_k \in \boldsymbol{\mathcal{Y}} \cap \boldsymbol{\mathcal{X}}, \quad y_k \xrightarrow{\boldsymbol{\mathcal{Y}}} \hat{y}, \quad y_k \xrightarrow{\boldsymbol{\mathcal{X}}} \hat{x} \Longrightarrow \hat{y} = \hat{x} \in \boldsymbol{\mathcal{Y}} \cap \boldsymbol{\mathcal{X}}.$$

Assume that, in addition to $(\mathbf{H}_1)$(i), *the linear operator* $\mathbf{L}$ *is continuous from* $\boldsymbol{\mathcal{V}}$ *to* $\boldsymbol{\mathcal{Y}}$ *and that the nonlinear map* $\mathbf{F}$ *is Lipschitzian from* $\boldsymbol{\mathcal{Y}}$ *to* $\boldsymbol{\mathcal{V}}$ *with Lipschitz constant* C *such that* $C\|\mathbf{L}\|_{\boldsymbol{\mathcal{V}}\to\boldsymbol{\mathcal{Y}}} =: \vartheta < 1$. *Then:*

(i) *One has* $(\mathbf{H}_1)$(ii) *with* $\mathbf{G}_F$ *Lipschitzian.*

(ii) *If* (*in addition to the original hypotheses*) *one has a growth condition*

$$|\mathbf{F}y|_{\boldsymbol{\mathcal{V}}} \leq a_0 + a_1|y|^{\bar{r}}_{\boldsymbol{\mathcal{Y}}} \tag{6.48}$$

for $\mathbf{F}$, *then* $\mathbf{G}_{\mathrm{F}}$ *satisfies the growth condition* (4.28).

(iii) *If* (*in addition to the original hypotheses*) *one has*

$$\{\mathbf{F}(\mathbf{L}g + z) : g \in \boldsymbol{\mathcal{G}}_0,\ z \in \boldsymbol{\mathcal{Z}}_0\} \text{ precompact in } \boldsymbol{\mathcal{V}} \quad \text{for each compact subset } \boldsymbol{\mathcal{G}}_0 \subset \boldsymbol{\mathcal{V}}, \tag{6.49}$$

then $\mathbf{G}_{\mathrm{F}}(\boldsymbol{\mathcal{Z}}_0)$ *is precompact in* $\boldsymbol{\mathcal{V}}$. (*Note that if* (6.49) *holds for each bounded subset* $\boldsymbol{\mathcal{Z}}_0$ *of* $\boldsymbol{\mathcal{Z}}$, *then* (iii) *gives* $(\mathbf{H}_1)$(iii); *a sufficient condition for this, in view of* (i), *is that the embedding* $\boldsymbol{\mathcal{Z}} \hookrightarrow \boldsymbol{\mathcal{Y}}$ *is compact.*)

Proof. Under the assumption $\vartheta < 1$, the map: $y \mapsto [\mathbf{LF}y + z]$ is a contraction on $\boldsymbol{\mathcal{Y}}$ so there is a unique fixpoint $y = Y(z)$, the solution of (3.15). One easily obtains the Lipschitz constant $1/(1-\vartheta)$ for the solution map $Y(\cdot) : z \mapsto y : \boldsymbol{\mathcal{Y}} \to \boldsymbol{\mathcal{Y}}$ and composing this with the Lipschitzian maps $\boldsymbol{\mathcal{Z}} \hookrightarrow \boldsymbol{\mathcal{Y}}$ on one side and $\mathbf{F}$ on the other gives the desired Lipschitz continuity for $\mathbf{G}_{\mathrm{F}} : z \mapsto \mathbf{F}y : \boldsymbol{\mathcal{Z}} \to \boldsymbol{\mathcal{V}}$. This gives linear growth for $Y(\cdot) : \boldsymbol{\mathcal{Z}} \to \boldsymbol{\mathcal{Y}}$ and composing that with (6.48) gives (4.28). The compactness assertion (iii) is an immediate consequence of Lemma 2 in [14], applied to the uniformly contractive family of maps: $g \mapsto \mathbf{F}(\mathbf{L}g + z)$ $(z \in \boldsymbol{\mathcal{Z}}_0)$ whose fixpoints give $\mathbf{G}_{\mathrm{F}}(\boldsymbol{\mathcal{Z}}_0)$. □

Note that our treatment of these problems presupposes well-posedness for the equations and, for our application of fixed-point arguments, convexity of $\boldsymbol{\mathcal{Z}}_{ad}$ and suitable compactness.

We now turn to a somewhat more general consideration of abstract settings giving $(\mathbf{H}_1)$. For this consideration our present concerns are essentially identical with those of [15] and we recall the relevant discussion there in providing classes of settings for which the relevant well-posedness and compactness can be verified. For this we suppose $\mathbf{A}(\cdot)$ generates a *fundamental solution* (evolution system) $\mathbf{S}$ — i.e.,

$$\begin{array}{lll} (6.50) & \text{(i)} & \mathbf{S}(t,s) \text{ is a bounded linear operator on } \mathcal{X} \text{ with} \\ & & \|\mathbf{S}(t,s)\| \leq M \text{ for } 0 \leq s \leq t \leq T; \\ & \text{(ii)} & \mathbf{S}(t,s)\mathbf{S}(s,r) = \mathbf{S}(t,r) \text{ for } 0 \leq r \leq s \leq t \leq T; \\ & \text{(iii)} & \mathbf{S}(t,s)\xi \longrightarrow \xi \text{ as } t \to s+ \text{ for } \xi \in \mathcal{X}; \\ & \text{(iv)} & \partial\mathbf{S}(t,s)\xi/\partial t = \mathbf{A}(s)\xi \text{ for } t = s \text{ and } \xi \in \mathcal{D} = \mathcal{D}(\mathbf{A}(s)). \end{array}$$

This permits us to introduce the notion of a *mild solution* [9] of (1.2) or (1.4): the linear map $\mathbf{L}$ will be given, in terms of $\mathbf{S}(\cdot)$, by

$$[\mathbf{L}v](t) = \mathbf{L}_t v := \int_0^t \mathbf{S}(t,s)v(s)\,ds \text{ for } 0 \leq t \leq T \tag{6.51}$$

for suitable $v(\cdot)$. In this formulation, (1.1) corresponds to the nonlinear integral equation[5] (abstract Volterra equation of second kind):

$$x(t) = \bar{z}(t) + \int_0^t \mathbf{S}(t,s)[f(s,x(s)) + [\mathbf{B}(u)](s)]\,ds$$

[5] Here $\bar{z}(t)$ is the solution of the equation $\dot{z} = \mathbf{A}z$ with the original initial conditions. With our assumption that the problem has been formulated so as to have homogeneous initial conditions, this term vanishes: $\bar{z} = 0$.

This is, of course, just the operator equation (3.15).

Now introduce (reflexive) Banach spaces $\mathcal{V}$, $\mathcal{W}$ and spaces $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{\mathcal{W}}$ of $\mathcal{V}$- and $\mathcal{W}$-valued functions, respectively, on $[0,T]$. We assume $\boldsymbol{\mathcal{V}}$, $\boldsymbol{\mathcal{W}}$ are compatible with $\mathcal{X}$ in the sense that the set $\mathcal{V} \cap \mathcal{X}$ is dense both in $\mathcal{V}$ and in $\mathcal{X}$ with

$$v_k \in \mathcal{V} \cap \mathcal{X}, \quad v_k \xrightarrow{\mathcal{V}} \hat{v}, \quad v_k \xrightarrow{\mathcal{X}} \hat{x} \Longrightarrow \hat{v} = \hat{x} \in \mathcal{V} \cap \mathcal{X}$$

and similarly for $[\mathcal{W}, \mathcal{X}]$. It will thus be possible to make suitable extensions or restrictions of $\mathbf{S}(t,s)$ so, e.g., the formal definition (3.14) may make sense for v in $\boldsymbol{\mathcal{V}}$ or in $\boldsymbol{\mathcal{W}}$.

We note that [15] provides four alternate sets of more concrete conditions on $\mathcal{X}, \mathcal{W}, \mathcal{V}, \mathbf{S}(\cdot\cdot), f(\cdot\cdot)$ under which one can verify $(\mathbf{H}_1)$. For convenience of reference we present these here, converted to our present notation. For this, we take $\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{W}}$ to have the form

$$\boldsymbol{\mathcal{V}} := L^p([0,T] \to \mathcal{V}), \qquad \boldsymbol{\mathcal{W}} := L^{p'}([0,T] \to \mathcal{W}) \tag{6.52}$$

and introduce another possible space $\mathcal{Y}$ compatible with $\mathcal{X}$. We assume:

$(\mathbf{C}_1)$ Let *each* of the following hold:

(i) $\|\mathbf{S}(t,s)\|_{\mathcal{V}\to\mathcal{X}} \le \rho_{\mathcal{V}}(t-s)$, $\|\mathbf{S}(t,s)\|_{\mathcal{W}\to\mathcal{X}} \le \rho_{\mathcal{W}}(t-s)$,
$\|\mathbf{S}(t,s)\|_{\mathcal{V}\to\mathcal{Y}} \le \hat{\rho}_{\mathcal{V}}(t-s)$, $\|\mathbf{S}(t,s)\|_{\mathcal{W}\to\mathcal{Y}} \le \hat{\rho}_{\mathcal{W}}(t-s)$;

(ii) $\|\mathbf{S}(t,s) - \mathbf{S}(t',s)\|_{\mathcal{Y}\to\mathcal{X}}$, $\|\mathbf{S}(t,s) - \mathbf{S}(t',s)\|_{\mathcal{W}\to\mathcal{X}} \le \varepsilon$
for $0 \le s \le t' - \varepsilon$, $t' < t \le T$ with $\varepsilon = \varepsilon(h) \to 0$ as $h := t - t' \to 0$;

(iii) $|f(t,\eta)|_{\mathcal{V}} \le \alpha(t) + \beta|\eta|_{\mathcal{Y}}^{\bar{r}}$ $\quad (\bar{r} := \bar{p}/p < 1)$
$|\mathbf{S}(t,s)[f(s,\eta) - f(s,\eta')]|_{\mathcal{Y}} \le \rho_{\mathcal{Y}}(t-s)|\eta - \eta'|_{\mathcal{Y}}$

where

$$\rho_{\mathcal{V}} \in L^q, \quad \rho_{\mathcal{W}} \in L^{q'} \quad \hat{\rho}_{\mathcal{V}} \in L^{\bar{q}}, \quad \hat{\rho}_{\mathcal{W}} \in L^{\bar{q}'}, \quad \alpha \in L^p, \quad \rho_{\mathcal{Y}} \in L^1,$$

with $1 < p$, $p' < \infty$, $1 \le \bar{p} < p$, and

$$1/p + 1/q = 1/p' + 1/q' = 1; \;\; 1/p + 1/\bar{q}, 1/p' + 1/\bar{q}' \le 1 + 1/\bar{p}.$$

To the set of conditions $(\mathbf{C}_1)$ we may adjoin any of four *alternative* conditions:

$(\mathbf{C}_2)$ Let *any one* of the following hold:

(i) For some Banach space $\hat{\mathcal{Y}} = \hat{\mathcal{Y}}_i$ such that the embedding: $\mathcal{Y} \hookrightarrow \hat{\mathcal{Y}}$ is compact, assume that for small $\delta > 0$ there exists M_δ such that
$\|\mathbf{S}(t, t-\delta)\|_{\hat{\mathcal{Y}}\to\mathcal{Y}} \le M_\delta \qquad$ for $\delta \le t \le T$.

(ii) For some Banach space $\hat{\mathcal{Y}} = \hat{\mathcal{Y}}_{ii}$ such that the embedding $\hat{\mathcal{Y}} \hookrightarrow \mathcal{Y}$ is compact, strengthen $(\mathbf{C}_1)$(i) by requiring:
$\|\mathbf{S}(t,s)\|_{\mathcal{W}\to\hat{\mathcal{Y}}} \le \hat{\rho}_{\mathcal{W}}(t-s) \qquad$ ($\hat{\mathcal{Y}}$ replacing $\mathcal{Y}$).

(iii) For some Banach space $\hat{\mathcal{Y}} = \hat{\mathcal{Y}}_{iii}$ such that the embedding: $\mathcal{Y} \hookrightarrow \hat{\mathcal{Y}}$ is compact, strengthen the growth condition in $(\mathbf{C}_1)$(iii) by requiring
$|f(t,\zeta)|_{\mathcal{V}} \le \alpha(t) + \beta|\zeta|_{\hat{\mathcal{Y}}}^{\bar{r}}$.
with $\bar{r} := \bar{p}/p < 1$ ($\hat{\mathcal{Y}}$ replacing $\mathcal{Y}$).

(iv) Take $\mathcal{Y} = \mathcal{X}$ reflexive in $(\mathbf{C}_1)$ and, for some Banach space $\hat{\mathcal{Y}} = \hat{\mathcal{Y}}_{iv}$ such that $\mathcal{X} = \mathcal{Y} \hookrightarrow \hat{\mathcal{Y}}$ is a compact embedding, assume that for each $\mu > 0$ there exists $\alpha_\mu \in L^p$ for which
$|\zeta|_{\hat{\mathcal{Y}}} \le \mu \Longrightarrow |f(t,\zeta)|_{\mathcal{V}} \le \alpha_\mu(t)$.

LEMMA 4. *Let $\mathcal{X}, \mathcal{W}, \mathcal{V}$, etc., be as above and assume* $(\mathbf{C}_1)$; *we will norm* $\boldsymbol{\mathcal{Z}} := \mathbf{L}\boldsymbol{\mathcal{W}}$ *so* $\mathbf{L}$ *is an isometry. Then one has* $(\mathbf{H}_1)$(i) *and* $(\mathbf{H}_1)$(ii) *as well as the growth condition* (4.28). *If in addition, we assume* $(\mathbf{C}_2)$ (*i.e., any one of the four alternative conditions presented*), *then* $\mathbf{G}$ *is also compact:*

$$(\mathbf{C}_1) + (\mathbf{C}_2) \Longrightarrow (H_1) + (4.28).$$

Proof. See [15] for details. We note here that only $(\mathbf{C}_1)$(ii) is used to give $(\mathbf{H}_1)$(i) and that $(\mathbf{C}_1)$(i),(iii) give a solution of (3.15) initially in $\boldsymbol{\mathcal{Y}} := L^{\bar{p}}([0,T] \to \mathcal{Y})$ from which one obtains $g := f(\cdot, x(\cdot)) \in \boldsymbol{\mathcal{V}}$ by Krasnoselskii's Theorem (cf., e.g., [2]). One then has $x \in \boldsymbol{\mathcal{X}}$ from $x = \mathbf{L}g + \mathbf{L}w$ and $(\mathbf{H}_1)$(i). The four alternative arguments for compactness of $\mathbf{G}$ from $(\mathbf{C}_2)$ use, among them, the Aubin Compactness Theorem [1], the Arzela–Ascoli Theorem, and an argument from [14]. □

Remark. The conditions above do not apply directly to boundary control (as in Examples 2 and 3) but the arguments in [15] may be easily modifiable to treat this situation. Indeed, the arguments in [15] are fairly standard, using convolution estimates from the form of $(\mathbf{C}_1)$(i) as applied to the representation (6.51) and the modification simply amounts to the corresponding use of a rather similar representation [3]

$$z(t) = [\mathbf{L}w](t) = \int_0^t [(-\mathbf{A})^{\vartheta}\mathbf{S}(t-s)][(-\mathbf{A})^{1-\vartheta}\Gamma]w(s)\,ds \tag{6.53}$$

for the solution of

$$\dot{z} = \mathbf{A}z, \qquad \mathbf{M}z = w := \mathbf{B}(u) \in \boldsymbol{\mathcal{W}}_{ad}. \tag{6.54}$$

Here, $\mathbf{A}$ (whose domain formally involves imposition of homogeneous boundary conditions) is the infinitesimal generator of an analytic semigroup $\mathbf{S}$ on $\mathcal{X}$ and $\mathbf{M}$ is the appropriate boundary operator; we have introduced the so-called "Green's operator" $\Gamma : \omega \mapsto v$ by

$$-\mathbf{A}v = 0, \qquad \mathbf{M}v = \omega.$$

We then use regularity theory for the problem defining Γ, the relation of the Sobolev spaces $H^{2\vartheta} := [L^2, H^1]_{2\vartheta}$ to the domain of $(-\mathbf{A})^{\vartheta}$ (e.g., [7] or [8]) when $\mathbf{A}$ is a second order elliptic operator, and the estimate (2.9).

We now turn to consideration of the convexity of $\boldsymbol{\mathcal{Z}}_{ad}$ and ask: How restrictive an assumption is this? We have already noted that we may replace $\boldsymbol{\mathcal{Z}}_{ad}$ by any *equivalent* set $\boldsymbol{\mathcal{Z}}'_{ad}$ satisfying (3.17) and we shall now see that we may quite reasonably expect to find such a closed convex set $\boldsymbol{\mathcal{Z}}'_{ad}$. We will actually work with $\boldsymbol{\mathcal{W}}_{ad}$ (cf. (3.18)) and will find a natural setting for which $\boldsymbol{\mathcal{W}}_{ad} \equiv \boldsymbol{\mathcal{W}}'_{ad} := \overline{co}\boldsymbol{\mathcal{W}}_{ad}$ ($\overline{co}$ = closed convex hull in $\boldsymbol{\mathcal{W}}$). We first need a preparatory result, which we "dignify" as a theorem.

THEOREM 7. *Let* $\boldsymbol{\mathcal{W}}$ *have the form* $L^p([0,T] \to \mathcal{W})$ $[1 \le p < \infty]$ *with* $\mathcal{W}$ *satisfying the technical hypothesis that*[6]

$$\boldsymbol{\mathcal{W}}_* := L^{\infty}([0,T] \to \mathcal{W}^*) \quad \textit{is dense in } \boldsymbol{\mathcal{W}}^*. \tag{6.55}$$

Suppose a subset $\boldsymbol{\mathcal{W}}_0 \subset \boldsymbol{\mathcal{W}}$ *has*[7] *the* segment property*:*

[6] It is known to be sufficient for this that $\mathcal{W}$ be reflexive or, somewhat more generally, that $\mathcal{W}$ have the Radon-Nikodym Property; cf., e.g., [6].

[7] Typically, $\boldsymbol{\mathcal{U}}_{ad}$ has the form: $\boldsymbol{\mathcal{U}}_{ad} = \{u \in \boldsymbol{\mathcal{U}} : u(t) \in \mathcal{U}_{ad}(t) \subset \mathcal{U} \text{ a.e.}\}$ and so satisfies (SP) if, say, $\boldsymbol{\mathcal{U}} := L^p([0,T] \to \mathcal{U})$. In this case, if $\mathbf{B}$ is defined pointwise in t for (1.2), etc., then (SP) is immediate for $\boldsymbol{\mathcal{W}}_{ad}$.

(SP) $\qquad w_0, w_1 \in \boldsymbol{\mathcal{W}}_0 \Longrightarrow w_s \in \boldsymbol{\mathcal{W}}_0 \quad (s \in \mathcal{S} = \{\textit{dense in } [0,T]\})$
where, given w_0 *and* w_1, *we set*

$$w_s(t) := \{w_1(t) \textit{ for } 0 \le t < s; \quad w_0(t) \textit{ for } s \le t \le T\}.$$

Then $\boldsymbol{\mathcal{W}}_1$, *the sequential weak closure of* $\boldsymbol{\mathcal{W}}_0$ *in* $\boldsymbol{\mathcal{W}}$, *is just the closed convex hull* $\overline{co}\boldsymbol{\mathcal{W}}_0$.

Proof. We set $\Sigma_{\mathcal{S}} :=$ {finite unions of intervals in $[0,T]$ with endpoints in $\mathcal{S}$}; for $\sigma \in \Sigma_{\mathcal{S}}$, let χ_σ be its characteristic function and then for $w_0, w_1 \in \boldsymbol{\mathcal{W}}_0$ set

$$w_\sigma = \{w_1 \text{ on } \sigma; \quad w_0 \text{ on } [0,T]\backslash\sigma\} = (1-\chi_\sigma)w_0 + \chi_\sigma w_1$$

so, e.g., w_s — as defined for (SP) — is now $w_{[0,s)}$. Note that repeated application of (SP) shows that each such w_σ must also be in $\boldsymbol{\mathcal{W}}_0$. Note, also, that (SP) for $\boldsymbol{\mathcal{W}}_0$ immediately gives (SP) for its sequential weak closure $\boldsymbol{\mathcal{W}}'$: just consider approximating sequences from $\boldsymbol{\mathcal{W}}_0$. Since $\overline{co}\boldsymbol{\mathcal{W}}_1 = \overline{co}\boldsymbol{\mathcal{W}}_0$, we may assume, with no loss of generality, that $\boldsymbol{\mathcal{W}}_0$ is already (sequentially weakly) closed and only show that (SP) then implies its convexity.

Now observe that for any constant $\vartheta \in (0,1)$ there is a sequence $\{\sigma_n\} \subset \Sigma_{\mathcal{S}}$ such that $\{\chi_\sigma\}$ is weak-* convergent to the constant function ϑ in $L^\infty(0,T)$. To see this, just partition $[0,T]$ into n equal subintervals and (noting the assumed density of $\mathcal{S}$) choose a sub-subinterval in each with endpoints in $\mathcal{S}$ and having length approximately $\vartheta T/n$; this gives σ_n. The weak-* convergence to ϑ is then an easy computation.

Given any w_0, $w_1 \in \boldsymbol{\mathcal{W}}_0$ and $\vartheta \in (0,1)$, we note, by repeated application of (SP), that each $w_\sigma = [\chi_\sigma w_1 + (1-\chi_\sigma)w_0]$ is also in $\boldsymbol{\mathcal{W}}_0$. For any $\psi \in \boldsymbol{\mathcal{W}}_*$ the $\mathcal{W}$-duality product $\langle \psi, w_1 - w_0\rangle$ (taken pointwise in t) is a function in $L^p(0,T) \subset L^1(0,T)$ so for the $\boldsymbol{\mathcal{W}}$-duality we have

$$\begin{aligned} \langle \psi, w_\sigma \rangle &= \int_0^T \langle \psi, \chi_\sigma w_1 + (1-\chi_\sigma)w_0\rangle\, dt \\ &= \int_0^T \chi_\sigma \langle \psi, w_1 - w_0\rangle\, dt + \langle \psi, w_0\rangle \\ &= \langle \chi_\sigma, \langle \psi, w_1 - w_0\rangle\rangle_{L^1(0,T)} + \langle \psi, w_0 \rangle \\ &\longrightarrow \langle \vartheta, \langle \psi, w_1 - w_0\rangle\rangle_{L^1(0,T)} + \langle \psi, w_0 \rangle \\ &= \langle \psi, \vartheta w_1 + (1-\vartheta)w_0\rangle. \end{aligned}$$

Since this holds for each ψ in the dense subset $\boldsymbol{\mathcal{W}}_* \subset \boldsymbol{\mathcal{W}}^*$, it follows that $w_\sigma \rightharpoonup [\vartheta w_1 + (1-\vartheta)w_0]$ (weak convergence in $\boldsymbol{\mathcal{W}}$) so this is in $\boldsymbol{\mathcal{W}}_0$. This, for each such w_0, w_1, ϑ, is just the desired convexity of $\boldsymbol{\mathcal{W}}_0$. □

COROLLARY. *Assume* $(\mathbf{H}_1)$. *Let* $\boldsymbol{\mathcal{W}}$, $\boldsymbol{\mathcal{W}}_{ad}$ *be as in Theorem* 7. *Suppose* $\mathbf{S} : w \mapsto z : \boldsymbol{\mathcal{W}} \to \boldsymbol{\mathcal{Z}}$, *as in* (3.18), *is continuous with* $\mathbf{ES} : \boldsymbol{\mathcal{W}} \to \mathcal{X}$ *compact and the graph of* $\mathbf{G}_{\mathrm{F}} \circ \mathbf{S}$ *is closed in* $\boldsymbol{\mathcal{W}}_{sw} \times \boldsymbol{\mathcal{V}}$ ($\cdot_{sw}$ *indicates the sequential weak topology*). *Then* $\boldsymbol{\mathcal{W}}_{ad} \equiv \boldsymbol{\mathcal{W}}'_{ad} := \overline{co}\boldsymbol{\mathcal{W}}_{ad}$ *in the sense of* (3.17).

Proof. Since $\boldsymbol{\mathcal{W}}_{ad} \subset \boldsymbol{\mathcal{W}}'_{ad}$, we automatically have similar inclusions in (3.17). It is the converse which is effectively a corollary to Theorem 7.

Fix $\xi \in \bar{\mathcal{K}}_{\mathrm{F}}(\boldsymbol{\mathcal{Z}}'_{ad})$ where $\boldsymbol{\mathcal{Z}}'_{ad} := \mathbf{S}[\overline{co}\boldsymbol{\mathcal{W}}_{ad}]$ so for each (fixed) $\varepsilon > 0$ there is some $w = w_\varepsilon \in \overline{co}\boldsymbol{\mathcal{W}}_{ad}$ such that $|\xi - \mathbf{TS}w| < \varepsilon$; set $z = \mathbf{S}w$ and $g = \mathbf{G}z$. By Theorem 7 there is then a sequence w_k in $\boldsymbol{\mathcal{W}}_{ad}$ with $w_k \rightharpoonup w$. Since $\{z_k := \mathbf{S}w_k\}$ is bounded we may, by $(\mathbf{H}_1)$(iii), extract a subsequence (still denoted by w_k) with $g_k = \mathbf{G}z_k \to \hat{g}$ in $\boldsymbol{\mathcal{V}}$. The closed graph requirement then ensures that $\hat{g} = g$ so, in particular, $\mathbf{L}_{[\mathrm{T}]}g_k \to \mathbf{L}_{[\mathrm{T}]}g$ in $\boldsymbol{\mathcal{X}}$. Since we also have $\mathbf{ES}w_k \to \mathbf{ES}w$ by the assumed compactness, this gives $\mathbf{TS}w_k \to \mathbf{TS}w$. This shows that for some $w_k \in \boldsymbol{\mathcal{W}}_{ad}$ we have $|\xi - \mathbf{TS}w_k| < 2\varepsilon$ and that, for each $\varepsilon > 0$, proves that we have $\xi \in \bar{\mathcal{K}}_{\mathrm{F}}(\boldsymbol{\mathcal{Z}}_{ad})$. Of course this applies for each $\xi \in \bar{\mathcal{K}}_{\mathrm{F}}(\boldsymbol{\mathcal{Z}}'_{ad})$. The arguments for $\bar{\mathcal{K}}_0$ and $\bar{\mathcal{K}}_g$ are similar but simpler. □

Note that, with only a small modification of the hypotheses here, one can show that the closures (say, in $\boldsymbol{\mathcal{X}}$) of the solution sets for (1.2), (1.4), (1.1) will be the same for $\boldsymbol{\mathcal{W}}_{ad}$ as for its (closed) convex hull $\boldsymbol{\mathcal{W}}'_{ad}$.

REFERENCES

[1] J. P. Aubin, *Un théorème de compacité*, CRAS de Paris, 1963, pp. 5042–5043.

[2] J. P. Aubin and I. Ekeland, *Applied Nonlinear Analysis.* Wiley-Interscience, New York, 1984.

[3] A. V. Balakrishnan, *Applied Functional Analysis.* Springer Verlag, New York, Berlin, 1976.

[4] J. Bergh and J. Löfstrom, *Interpolation Spaces, an Introduction.* Springer-Verlag, Berlin, 1976.

[5] N. Carmichael and M.D. Quinn, *Fixed point methods in nonlinear control*, in Distributed Parameter Systems, Lecture Notes in Control and Information Sci. 75; F. Kappel, K. Kunisch, and W. Schappacher, eds., Springer-Verlag, New York, Berlin, 1985, pp. 24–51.

[6] J. Diestel and J. J. Uhl, *Vector Measures.* American Mathematical Society, Providence, RI, 1977.

[7] D. Fujiwara, *Concrete characterization of the domains of fractional powers of some elliptic differential operators of the second order*, Proc. Japan Acad., 43 (1967), pp. 82–86.

[8] P. Grisvard, *Caractérisation de quelques espaces d'interpolation*, Arch. Rat. Mech. Anal., 25 (1967), pp. 40–63.

[9] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*; Lecture Notes in Math. 840, Springer-Verlag, Berlin, 1981.

[10] E. Michael, *Continuous selections* I, Ann. Math., 63 (1956), pp. 361–382.

[11] K. Naito, *Controllability of semilinear control systems, I*, SIAM J. Control Optim., 25 (1987), pp. 715–722.

[12] K. Naito, *Approximation and controllability for solutions of semilinear control systems*, Control-Theory and Advanced Technology 1 (1985), pp. 165–173.

[13] T. I. Seidman, *A one-sided miscellany*, in Proc. IFAC 4th Symposium on Distributed Parameter Systems, 1986.

[14] T. I. Seidman, *Two compactness lemmas*, in Nonlinear Semigroups, Partial Differential Equations, and Attractors; Lecture Notes in Math. 1248, T. Gill and W. Zachary, eds., Springer-Verlag, New York, Berlin, 1987, pp. 162–168.

[15] T. I. Seidman, *Invariance of the reachable set under nonlinear perturbations*, SIAM J. Control Optim., 25 (1987), pp. 1173-1191.

[16] T. I. Seidman, *Invariance under nonlinear perturbations for reachable and almost-reachable sets*, in Control Problems for Systems Described by Partial Differential Equations and Applications, Lecture Notes Control and Information Sci. 97, I. Lasiecka and R. Triggiani, eds., Springer-Verlag, New York, Berlin, 1987, pp. 336–345.

[17] E. Zuazua, *Exact controllability for the semilinear wave equation*, in Nonlinear Partial Differential Equations and Their Applications, Collége de France Seminar '87–'88, H. Brezis and J. L. Lions, eds., Pitman, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 751-768, July 1991

SOME RESULTS FOR AN OPTIMAL CONTROL PROBLEM WITH SEMILINEAR STATE EQUATION*

FAUSTO GOZZI†

Abstract. A control problem governed by a semilinear state eqution depending on a small parameter ε is considered. It is proven that, for ε sufficiently small, the associated Hamilton-Jacobi equation has a unique strict solution; consequently, the control problem can be solved by dynamic programming method.

Key words. optimal control, semilinear state equation, Hamilton-Jacobi equation

AMS(MOS) subject classifications. 93C, 49B

1. Introduction. We are concerned with a dynamical system governed by the following semilinear state equation:

$$\text{(SE)}\qquad \begin{aligned} &y' = Ay + f(\varepsilon, y) + u \quad \text{on } [0, T],\\ &y(0) = x, \qquad x \in H, \end{aligned}$$

where $A: D(A) \subset H \to H$ is the infinitesimal generator of a strongly continuous semigroup in a Hilbert space H, ε is a real parameter, and f is a smooth function from H to H which goes uniformly to zero (with its derivatives) when $\varepsilon \to 0$; T is a fixed positive constant.

We consider then the following optimal control problem (P):

Minimize the functional

$$\text{(P)}\quad (1.1)\qquad J(x, u) = \int_0^T \left[g(y(s)) + \frac{1}{2}|u(s)|^2 \right] ds + \phi_0(y(T))$$

over all controls $u \in L^2(0, T; U)$,
where y is subject to the state equation (SE).

Here g and ϕ_0 are smooth convex functions from H to H.

We treat this problem by the dynamic programming method studying the corresponding Hamilton-Jacobi equation:

$$\text{(HJ)}\qquad \begin{aligned} &\phi_t(t, x) - \tfrac{1}{2}|\phi_x(t, x)|^2 + \langle Ax + f(\varepsilon, x), \phi_x(t, x)\rangle + g(x) = 0 \\ &\qquad\qquad \forall (t, x) \in [0, T] \times D(A),\\ &\phi(T, x) = \phi_0(x) \quad \forall x \in H. \end{aligned}$$

If we set $\varepsilon = 0$, $g(x) = \frac{1}{2}\langle Mx, x\rangle$, and $\phi_0(x) = \frac{1}{2}\langle P_0 x, x\rangle$, then the problem (P) reduces to the well-known linear regulator problem, which has been extensively studied (see, for instance, [7]). However, for the applications, the use of a linear state equation is very restrictive.

Moreover, if g and ϕ_0 are general convex functions and $\varepsilon = 0$, the cost J is a convex function on $L^2(0, T; H)$ and (HJ) admits a unique strict solution on $[0, T]$ (see [2] and [3]). We remark that if $\varepsilon \neq 0$ then J is generally not convex and the method used in [4] only gives the existence of a local solution of (HJ) in some interval $[0, \delta]$.

* Received by the editors April 20, 1988; accepted for publication (in revised form) August 9, 1990.

† Dipartimento di Matematica, Scuola Normale Superiore, Via F. Buonarroti, 56127 Pisa, Italy.

The viscosity solutions approach has recently been generalized by Crandall and Lions (see [5]) to cover infinite-dimensional problems. They can prove, under suitable hypotheses, existence and uniqueness of a viscosity solution ϕ of (HJ). However, ϕ is not regular, and the classical method of dynamic programming based on the solution of the closed-loop equation,

$$\text{(1.2)}\qquad \begin{aligned} &y'(s) = Ay(s) + f(\varepsilon, y(s)) - \phi_x(T-s, y(s)) \quad \text{on } [0, T],\\ &y(0) = x, \qquad x \in H, \end{aligned}$$

cannot be used.

When the semigroup $\{e^{tA}\}_{t \geqq 0}$ is compact it is possible to study (HJ) by variational methods (see [1], for instance). In this case the existence of a feedback optimal control can be proved; however, the solution ϕ is still not regular and the closed-loop equation (1.2) cannot be directly solved. For other results in this direction see [9].

The aim of this paper is to find, for $|\varepsilon|$ sufficiently small, global regular solutions of the Hamilton-Jacobi equation (HJ) *and then to solve the control problem* (P) *by* using the classical argument of dynamic programming.

More precisely, under suitable hypotheses, for every given $R > 0$ we prove that there exists $\bar{\varepsilon}(R) > 0$ such that for $|\varepsilon| \leqq \bar{\varepsilon}(R)$ equation (HJ) has a unique solution ϕ_ε in $[0, T] \times \Sigma_R$, where Σ_R is the closed ball in H with radius R.

Then we fix an upper bound r_0 for the norm of the initial state x, and we can show that, if ε is sufficiently small, the closed-loop equation (1.2) has a unique strong solution y_ε with $\sup_{t \in [0,T]} |y_\varepsilon(t)| \leqq Cr_0$ for a constant $C \geqq 1$ and that there exists a unique optimal control u_ε such that $\sup_{t \in [0,T]} |u_\varepsilon(t)| \leqq M(r_0)$ where M is a given function $(0, +\infty) \to (0, +\infty)$ depending only on the data. Moreover, we prove that the optimal pair $(u_\varepsilon, y_\varepsilon)$ is a regular function of the parameter ε and we give an expansion in a neighbourhood of $\varepsilon = 0$ in a special case.

Our hypotheses are that g and ϕ_0 are convex and sufficiently regular (two times differentiable), g is strictly convex, and $A - \omega \mathrm{I}$ is strictly dissipative for some $\omega > 0$.

We sketch briefly the idea of the proof of the main theorem. We consider the approximating equation:

$$\text{(HJ)}_\alpha \qquad \begin{aligned} &\phi_t(t, x) = -\frac{1}{\alpha}(\phi(t, x) - \phi_\alpha(t, x)) + \langle Ax + f(\varepsilon, x), \phi_x(t, x)\rangle + g(x)\\ &\hspace{20em} \forall (t, x) \in [0, T] \times D(A),\\ &\phi(0, x) = \phi_0(x) \quad \forall x \in H, \end{aligned}$$

where ϕ_α is the convex regularization of ϕ (see next section). Here the "bad" term $|\phi_x(t, x)|^2/2$ is replaced by $(\phi(t, x) - \phi_\alpha(t, x))/\alpha$ (see § 2).

When $\varepsilon = 0$ the solution ϕ_ε^α of (HJ)_α is convex and it is possible to show that ϕ_ε^α converges to the solution of (HJ) as $\alpha \to 0$ (see [2] and [3]). We are able to prove that if ε is sufficiently small, then the solution ϕ_ε^α still exists globally and it is convex. Thus it is possible to show the existence of the limit as $\alpha \to 0$.

2. Notation and preliminary results. We begin by specifying the notation we will use throughout this paper. Let H be a real Hilbert space with norm $|\cdot|_H$ and scalar product $\langle \cdot, \cdot \rangle_H$. With obvious modifications all the results of this paper can be adapted to complex Hilbert spaces. We denote by $\mathscr{L}(H)$ the Banach algebra of the linear bounded operators from H into H. By $\Sigma(H)$ we denote the set of all Hermitian operators in $\mathscr{L}(H)$, and we set

$$\Sigma^+ = \{T \in \mathscr{L}(H); \langle Tx, x \rangle \geqq 0, \forall x \in H\}.$$

Now let X, Y be two Hilbert spaces. If $f: X \to Y$ is a k-times Fréchet differentiable function, we set, for every $r > 0$,

$$|f|_{h,r} = \sup \{|f^{(h)}(x)|: |x| \leqq r\}, \qquad h = 0, \cdots, k \tag{2.1}$$

$$\|f\|_{h,r} = \sup \left\{ \frac{|f^{(h)}(x) - f^{(h)}(y)|}{|x-y|}: |x|, |y| \leqq r, x \neq y \right\}, \qquad h = 0, \cdots, k. \tag{2.2}$$

Note that if f is an $(n+1)$-times Fréchet differentiable function we have

$$\|f\|_{n,r} = |f|_{n+1,r}. \tag{2.3}$$

For $k \in \mathbf{N}$ we define

$$C^k(X, Y) = \{f: X \to Y, k\text{-times continuously Fréchet differentiable: } |f|_{h,r} < +\infty, \forall h = 0, \cdots, k, \forall r > 0\}, \tag{2.4}$$

$$C^k_{\text{Lip}}(X, Y) = \{f \in C^k(X, Y): \|f\|_{k,r} < +\infty, \forall r > 0\}. \tag{2.5}$$

If $Y = \mathbf{R}$ we write $C^k(X)$ instead of $C^k(X, \mathbf{R})$.

We remark that $C^k(X, Y)$ and $C^k_{\text{Lip}}(X, Y)$ are Fréchet spaces with the seminorms $\sum_{h=0}^k |f|_{h,r}$ and $\|f\|_{k,r} + \sum_{h=0}^k |f|_{h,r}$ $(r \in \mathbf{N})$, respectively, and we say that $f_n \to f$ in $C^k(X, Y)$ if $|f_n - f|_{i,r} \to 0$ for $i = 0, \cdots, k$ and for every $r > 0$.

Now let $\phi: [0, T] \times H \to \mathbf{R}$ and $k \geqq 1$. We say that $\phi \in B([0, T]; C^k(H))$ if ϕ satisfies the following conditions:

(i) $\sup_{t \in [0,T]} |\phi(t, \cdot)|_{h,r} < +\infty$, for $h = 0, \cdots, k$, for all $r > 0$,

(ii) $\phi: [0, T] \times H \to \mathbf{R}$ is continuous,

(iii) $\phi_x: [0, T] \times H \to H$ is continuous;

and, if $k \geqq 2$,

(iv) $\partial^h \phi / \partial x^h$ is strongly continuous for $h = 2, \cdots, k$, that is, the map

$$[0, T] \times H \to H, \qquad (t, x) \to \frac{\partial^h \phi}{\partial x^h}(y_1, \cdots, y_{h-1})$$

is continous for all $(y_1, \cdots, y_{h-1}) \in H^{h-1}$.

Analogously we say that $\phi \in B([0, T]; C^k_{\text{Lip}}(H))$ if $\phi \in B([0, T]; C^k(H))$ and

$$\sup_{t \in [0,T]} \|\phi(t, \cdot)\|_{k,r} < +\infty \quad \forall r > 0.$$

Moreover, we say that $\phi_n \to \phi$ in $B([0, T]; C^k(H))$ if

$$\sup_{t \in [0,T]} |\phi(t, \cdot) - \phi_n(t, \cdot)|_{i,r} \to 0 \quad \text{for } i = 0, \cdots, k, \quad \forall r > 0. \tag{2.6}$$

Finally, setting $\Sigma_r = \{x \in H, |x| \leqq r\}$, we similarly define the spaces $C^k(\Sigma_r)$, $C^k_{\text{Lip}}(\Sigma_r)$, $B([0, T]; C^k(\Sigma_r))$, $B([0, T]; C^k_{\text{Lip}}(\Sigma_r))$, and we set $|\phi|_{C^k(\Sigma_r)} = \sum_{i=0}^k |\phi|_{i,r}$, and $|\phi|_{C^k_{\text{Lip}}(\Sigma_r)} = |\phi|_{C^k(\Sigma_r)} + \|\phi\|_{k,r}$, for $\phi \in C^k(\Sigma_r)$ and $\phi \in C^k_{\text{Lip}}(\Sigma_r)$, respectively.

Now we consider the regularization of a convex function (see [2, p. 5]) and we recall its fundamental properties, which we will use in the following.

We denote by K (respectively, K_R) the set of all convex functions $\phi \in C^1(H)$ (respectively, $C^1(\Sigma_R)$) such that $\phi'(0) = 0$.

For $\phi \in K$ we set

$$\phi_\alpha(x) = \min \left\{ \phi(y) + \frac{1}{2\alpha} |x-y|^2; y \in H \right\}, \qquad \alpha \in \mathbf{R} - \{0\}, \quad x \in H. \tag{2.7}$$

We remark that the minimum exists and it is unique, due to the convexity of ϕ.

Moreover, setting

$$x_\alpha = (1+\alpha\phi')^{-1}(x), \qquad \alpha \in \mathbf{R}, \quad x \in H \tag{2.8}$$

we have

$$\phi_\alpha(x) = \phi(x_\alpha) + \frac{\alpha}{2}|\phi'(x_\alpha)|^2. \tag{2.9}$$

We now collect the properties of ϕ_α we need, which are proved in [2, pp. 35, 42].

LEMMA 2.1. *Let ϕ, $\bar\phi \in K \cap C^2_{\mathrm{Lip}}(H)$; then ϕ_α, $\bar\phi_\alpha \in K \cap C^2_{\mathrm{Lip}}(H)$ and the following estimates hold for every $R > 0$:*

(a) $|\phi_\alpha|_{i,R} \leqq |\phi|_{i,R}$, *for* $i = 0, 1, 2$;

(b) $\|\phi_\alpha\|_{2,R} \leqq \|\phi\|_{2,R}$;

(c) $|\phi_\alpha - \bar\phi_\alpha|_{i,R} \leqq |\phi - \bar\phi|_{i,R}$, *for* $i = 0, 1$;

(d) $|\phi_\alpha - \bar\phi_\alpha|_{2,R} \leqq |\phi - \bar\phi|_{2,R} + \alpha\|\bar\phi\|_{2,R}|\phi - \bar\phi|_{1,R}$;

LEMMA 2.2. *Let $\phi \in K \cap C^2_{\mathrm{Lip}}(H)$, and set*

$$R_{\alpha,\phi}(x) = \frac{1}{\alpha}(\phi(x) - \phi_\alpha(x)) - \frac{1}{2}|\phi'(x)|^2. \tag{2.10}$$

Then for $R > 0$ we have

(i) $|R_{\alpha,\phi}|_{0,R} \leqq \alpha|\phi|^2_{1,R}|\phi|_{2,R}$;

(ii) $|R_{\alpha,\phi}|_{1,R} \leqq \alpha(|\phi|^2_{1,R}\|\phi\|_{2,R} + |\phi|_{2,R}|\phi|_{1,R})$.

We observe that the same statements hold if K is replaced by K_R.

Remark 2.3. It is not difficult to show that if ϕ, $\bar\phi \in K \cap C^k_{\mathrm{Lip}}(H)$ ($k \geqq 3$), then analogous estimates hold true for $|\phi_\alpha|_{i,R}$, $|\phi_\alpha - \bar\phi_\alpha|_{i,R}$, and $|R_{\alpha,\phi}|_{i-1,R}$, $i = 3, \cdots, k$.

3. Global existence and uniqueness for the Hamilton–Jacobi equation. Consider the following Cauchy problem:

$$\text{(HJ)} \quad \begin{aligned} &\phi_t(t,x) = -\tfrac{1}{2}|\phi_x(t,x)|^2 + \langle Ax + f(\varepsilon,x), \phi_x(t,x)\rangle + g(x) \quad \forall (t,x) \in [0,T] \times D(A), \\ &\phi(0,x) = \phi_0(x) \quad \forall x \in H. \end{aligned}$$

Our hypotheses (as we said in the Introduction) are the following:

(i) $A: D(A) \subset H \to H$ is the infinitesimal generator of a strongly continuous semi-group in H and there exists $w \in \mathbf{R}$ subject to

$$\langle Ax, x\rangle \leqq \omega|x|^2 \quad \forall x \in D(A).$$

(ii) $f(\varepsilon, \cdot) \in C^2_{\mathrm{Lip}}(H, H)$, for all $\varepsilon \in \mathbf{R}$; $f(\varepsilon, 0) = 0$, $(\partial f/\partial x) f(\varepsilon, 0) = 0$, for all $\varepsilon \in \mathbf{R}$; $f(\varepsilon, x) \to_{\varepsilon \to 0} 0$ uniformly on $C^2_{\mathrm{Lip}}(H, H)$.

(iii) g, $\phi_0 \in C^2_{\mathrm{Lip}}(H) \cap K$; $\langle g''(x)z, z\rangle \geqq \mu|z|^2$, for all x, $z \in H$, for $\mu > 0$ fixed; that is, g is strictly convex on H.

We recall this result about (SE).

PROPOSITION 3.1. *Let $x \in H$, $u \in L^2(0, T; H)$; let $\bar\omega = \max\{\omega, 0\}$. Assume that* (i)–(iii) *hold true. We define*

$$r = 2\,e^{\bar\omega T}(|x|_H + \sqrt{T}\,|u|_{L^2(0,T;H)}) = r(|x|, |u|).$$

Then there exists $\varepsilon_0(r) > 0$ such that, for $|\varepsilon| \leqq \varepsilon_0$, (SE) *has a unique mild solution $y \in C([0, T]; H)$, and we have*

$$|y(t)| \leqq r \qquad \textit{and, if } \omega > 0 \qquad |y(t)| \leqq |x| + \sqrt{T}\,|u|_{L^2(0,T;H)}. \tag{3.1}$$

Sketch of proof. It is a standard application of contraction principle and Gronwall lemma (for similar arguments see [6] and [8, Chapter 6]). We consider (SE) in the following integral form:

$$\text{(SEI)} \qquad y(t)=e^{tA}x+\int_0^t e^{(t-s)A}[u(s)+f(\varepsilon, y(s))]\,ds \overset{\text{def}}{=} (\Lambda y)(t).$$

Let $B_r=\{y\in C([0,T];H);\ |y|\leqq r\}$. By simple estimates it follows that for ε sufficiently small, Λ defined in (SEI) is a contraction map on B_r. So there exists a unique solution of (SE) on B_r. From the Gronwall inequality we easily obtain that the solution y is unique also on $C([0,T];H)$ and that the estimates (3.1) hold. □

In the following for every fixed $r>0$ we will suppose that $|\varepsilon|\leqq\varepsilon_0(r)$.

We say that a function $\phi\in B([0,T];C^1_{\text{Lip}}(\Sigma_R)\cap K_R)$ is a strict solution of (HJ) if $\phi(\cdot,x)\in C^1([0,T])$ for all $x\in D(A)$, and satisfy (HJ).

More generally, if ϕ is defined only on $[0,T]\times\Sigma_R$ we say that ϕ is a strict solution on Σ_R if the conditions above hold true on this ball.

THEOREM 3.2. *Under assumptions* (i)–(iii), *for every* $R>0$ *there exists* $\bar\varepsilon(R)>0$ *such that, if* $|\varepsilon|\leqq\bar\varepsilon(R)$, *then the Hamilton–Jacobi equation* (HJ) *has a unique strict solution* $\phi_\varepsilon:[0,T]\times\Sigma_R\to\mathbf{R}$ *and we have*

$$\phi_\varepsilon(t,\cdot)\in C^2_{\text{Lip}}(\Sigma_R)\cap K_R \quad \forall t\in[0,T],$$

$$\sup_{t\in[0,T]} |\phi_\varepsilon(t,\cdot)|_{C^2_{\text{Lip}}(\Sigma_R)}<+\infty.$$

4. Proof of Theorem 3.2. First we prove, by the method of successive approximations, the existence of a solution ϕ^α_ε for the integral form of the approximating problem $(\text{HJ})_\alpha$. Afterwards we show that ϕ^α_ε converges to a strict solution ϕ_ε of the Cauchy problem (HJ) for $\alpha\to 0$.

The uniqueness is easily shown by standard methods.

At the moment we assume that hypothesis (i) holds with $\omega<0$, that is, A is a strictly dissipative operator on H. We will explain later (see Remark 4.8) the generalization of the proof to the case where $\omega\geqq 0$.

4.1. The integral approximating problem. Using the characteristics method we can write the approximating problem $(\text{HJ})_\alpha$ in the following integral form (see [4]):

$$\text{(4.1)} \qquad \begin{aligned}\phi(t,x)&=e^{-t/\alpha}\phi_0(\zeta_\varepsilon(0,t,x))\\ &\quad+\int_0^t e^{-(t-s)/\alpha}\left[g(\zeta_\varepsilon(s,t,x))+\frac{1}{\alpha}\phi_\alpha(s,\zeta_\varepsilon(s,t,x))\right]ds\end{aligned}$$

where $\zeta_\varepsilon(s,t,x)$ is the solution of the Cauchy problem:

$$\text{(4.2)} \qquad \begin{aligned}&\zeta'_\varepsilon(s)=-A\zeta_\varepsilon(s)-f(\varepsilon,\zeta_\varepsilon(s)) \quad \text{for } s\in[0,t],\\ &\zeta_\varepsilon(t)=x, \qquad x\in H, \quad t\in[0,T],\end{aligned}$$

that is, ζ_ε is the family of characteristic curves for $(\text{HJ})_\alpha$ linearized.

4.2. Characteristics estimate. We show some fundamental estimates for the characteristic curves.

PROPOSITION 4.1. *Let* (i), (ii) *be true. We fix* $R>0$. *Let* $t\in[0,T]$, $x\in H$ *with* $|x|\leqq R$. *There exists* $\varepsilon_1(R)>0$ *such that, if* $|\varepsilon|\leqq\varepsilon_1(R)$, *then the Cauchy problem* (4.2) *has a unique mild solution* $\zeta_\varepsilon(\cdot, t, x)$ *on* $[0, t]$ *and we have*

$$\zeta_\varepsilon(s, t, \cdot) \in C^2_{\mathrm{Lip}}(H, H) \quad \forall s, t \in [0, T], \quad s \leqq t.$$

Moreover, the following estimates hold:

$$|\zeta_\varepsilon(s, t, x)| \leqq |x|\, e^{+\omega(t-s)/2}, \tag{4.3}$$

$$|\zeta_\varepsilon(s, t, \cdot)|_{1,R} \leqq e^{+\omega(t-s)/2}, \tag{4.4}$$

$$|\zeta_{\varepsilon xx}(s, t, x)(z, z)|_H \leqq |f(\varepsilon, \cdot)|_{2,R} \int_s^t |\zeta_{\varepsilon x}(\sigma, t, x)z|_H^2\, e^{+\omega(\sigma-s)/2}\, d\sigma, \tag{4.5}$$

$$\|\zeta_\varepsilon(s, t, \cdot)\|_{2,R} \leqq \int_s^t e^{+\omega(\sigma-s)/2}[3|f(\varepsilon, \cdot)|_{2,R} + \|f(\varepsilon, \cdot)\|_{2,R}]|\zeta_\varepsilon(\sigma, t, \cdot)|_{1,R}\, d\sigma. \tag{4.6}$$

Proof. The proof of existence and uniqueness follows from Proposition 3.1. The regularity with respect to x follows from the parameter-dependent contraction principle. For the estimates it is sufficient to choose $\varepsilon_1(R)$ such that

$$\langle Ax + f(\varepsilon, x), x\rangle \leqq \frac{\omega}{2}|x|^2 \quad \forall x \in \Sigma_{2R} \quad \forall |\varepsilon| \leqq \varepsilon_1(R), \tag{4.7}$$

and it is possible due to hypothesis (ii). Now by (4.7) the operator $A + f(\varepsilon, \cdot)$ is strictly dissipative on Σ_{2R}.

If $|x| \leqq R$, then $|\zeta_\varepsilon(s, t, x)| \leqq 2R$ (see Proposition 3.1), and by applying Gronwall's inequality we obtain (4.3).

Analogously, $\zeta_{\varepsilon x}$ satisfies the Cauchy problem:

$$\begin{aligned} &\zeta'_{\varepsilon x}(s) = -A\zeta_{\varepsilon x}(x) - f_x(\varepsilon, \zeta_\varepsilon(s))\zeta_{\varepsilon x}(s) \quad \text{for } s \in [0, T], \\ &\zeta_{\varepsilon x}(s) = I, \qquad I \in \mathscr{L}(H) \end{aligned} \tag{4.8}$$

and again

$$\langle (A + f_x(\varepsilon, x))z, z\rangle \leqq \frac{\omega}{2}|z|^2 \quad \forall x \in \Sigma_R, \quad \forall z \in H.$$

From Gronwall's inequality (4.4) follows.

For the other estimates we observe that for every $z \in H$ we have

$$\begin{aligned} &\zeta'_{\varepsilon xx}(s)(z, z) = -A\zeta_{\varepsilon xx}(s)(z, z) - [f_x(\varepsilon, \zeta_\varepsilon(s))\zeta_{\varepsilon xx}(s)(z, z) \\ &\qquad\qquad + f_{xx}(\varepsilon, \zeta_\varepsilon(s))(\zeta_{\varepsilon x}(s)z, \zeta_{\varepsilon x}(s)z)] \quad \text{for } s \in [0, T], \\ &\xi_{\varepsilon xx}(t) = 0 \end{aligned} \tag{4.9}$$

and from Gronwall's inequality (4.5) and (4.6) follow.

4.3. Convexity estimates. We now consider the application

$$\Gamma^\varepsilon_\alpha : B([0, T]; C^2_{\mathrm{Lip}}(\Sigma_R)) \to B([0, T]; C^2_{\mathrm{Lip}}(\Sigma_R)) \tag{4.10}$$

defined by (see (4.1))

$$\begin{aligned} \Gamma^\varepsilon_\alpha \phi(t, x) &= e^{-t/\alpha}\phi_0(\zeta_\varepsilon(0, t, x)) \\ &\quad + \int_0^t e^{-(t-s)/\alpha}\left[g(\zeta_\varepsilon(s, t, x)) + \frac{1}{\alpha}\,\phi_\alpha(s, \zeta_\varepsilon(s, t, x))\right] ds. \end{aligned} \tag{4.11}$$

Clearly, $\Gamma^\varepsilon_\alpha$ is well defined and (4.10) holds. We prove that, for ε, α sufficiently small, $\Gamma^\varepsilon_\alpha$ is a convexity-preserving map.

LEMMA 4.2. *Let $R>0$, $C>0$ be fixed, and $\alpha\in[0,1]$. Let* (i)-(iii) *hold true. We consider*

$$\mathscr{F}_C^R=\left\{\phi\in B([0,T]);\ C^2_{\mathrm{Lip}}(\Sigma_R)\cap K_R)\ \textit{such that}\ \sup_{t\in[0,T]}|\phi(t,\cdot)|_{1,R}\leqq C\right\}.$$

Then there exists $\varepsilon(R,C)>0$ such that, for $|\varepsilon|\leqq\varepsilon(R,C)$, $\Gamma_\alpha^\varepsilon$ is a convexity-preserving map, that is,

$$\Gamma_\alpha^\varepsilon(\mathscr{F}_C^R)\subset B([0,T];\ C^2_{\mathrm{Lip}}(\Sigma_R)\cap K_R).$$

Proof. First we remark that, for $\phi\in C^2_{\mathrm{Lip}}(\Sigma_R)$, we have $\phi\in K_R$ if and only if

$$\langle\phi''(x)z,z\rangle\geqq 0\quad\forall x\in\Sigma_R,\quad\forall z\in H,\quad\phi'(0)=0.\tag{4.12}$$

Therefore we have to show that there exists $\varepsilon(R,C)>0$ such that for $|\varepsilon|\leqq\varepsilon(R,C)$

$$\langle(\Gamma_\alpha^\varepsilon\phi)_{xx}(t,x)z,z\rangle\geqq 0\quad\forall\phi\in\mathscr{F}_C^R,\quad\forall x\in\Sigma_R,\quad\forall z\in H,\quad\text{and}$$
$$(\Gamma_\alpha^\varepsilon\phi)_x(t,0)=0\quad\forall t\in[0,T].$$

Let $\phi\in B([0,T];\ C^2_{\mathrm{Lip}}(\Sigma_R))$ be such that $\sup_{t\in[0,T]}|\phi(t,\cdot)|_{1,R}\leqq C$. By differentiating twice (4.11) with respect to x we obtain

$$\begin{aligned}(\Gamma_\alpha^\varepsilon\phi)_x(t,x)&=e^{-t/\alpha}\zeta_{\varepsilon x}^*(0,t,x)\phi_0'(\zeta_\varepsilon(0,t,x))\\&\quad+\int_0^t e^{-(t-s)/\alpha}\zeta_{\varepsilon x}^*(s,t,x)\left[g'(\zeta_\varepsilon(s,t,x))+\frac{1}{\alpha}\phi_{\alpha x}(s,\zeta_\varepsilon(s,t,x))\right]ds,\end{aligned}\tag{4.13}$$

$$\begin{aligned}(\Gamma_\alpha^\varepsilon\phi)_{xx}(t,x)&=e^{-t/\alpha}\zeta_{\varepsilon x}^*(0,t,x)\phi_0''(\zeta_\varepsilon(0,t,x))\zeta_{\varepsilon x}(0,t,x)\\&\quad+\int_0^t e^{-(t-s)/\alpha}\zeta_{\varepsilon x}^*(s,t,x)\\&\quad\cdot\left[g''(\zeta_\varepsilon(s,t,x))+\frac{1}{\alpha}\phi_{\alpha xx}(s,\zeta_\varepsilon(s,t,x))\right]\zeta_{\varepsilon x}(s,t,x)\,ds\\&\quad+e^{-t/\alpha}\zeta_{\varepsilon xx}^*(0,t,x)\phi_0'(\zeta_\varepsilon(0,t,x))\\&\quad+\int_0^t e^{-(t-s)/\alpha}\zeta_{\varepsilon xx}^*(s,t,x)\left[g'(\zeta_\varepsilon(s,t,x))+\frac{1}{\alpha}\phi_{\alpha x}(s,\zeta_\varepsilon(s,t,x))\right]ds,\end{aligned}\tag{4.14}$$

and therefore

$$\begin{aligned}\langle(\Gamma_\alpha^\varepsilon\phi)_{xx}(t,x)z,z\rangle&=e^{-t/\alpha}\langle\phi_0''(\zeta_\varepsilon(0,t,x))\zeta_{\varepsilon x}(0,t,x)z,\zeta_{\varepsilon x}(0,t,x)z\rangle\\&\quad+\int_0^t e^{-(t-s)/\alpha}\left\langle\left[g''(\zeta_\varepsilon(s,t,x))+\frac{1}{\alpha}\phi_{\alpha xx}(s,\zeta_\varepsilon(s,t,x))\right]\right.\\&\qquad\left.\zeta_{\varepsilon x}(s,t,x)z,\zeta_{\varepsilon x}(s,t,x)z\right\rangle ds\\&\quad+e^{-t/\alpha}\langle\zeta_{\varepsilon xx}(0,t,x)(z,z),\phi_0'(\zeta_\varepsilon(0,t,x))\rangle+\int_0^t e^{-(t-s)/\alpha}\\&\quad\cdot\left\langle\zeta_{\varepsilon xx}(s,t,x)(z,z),\left[g'(\zeta_\varepsilon(s,t,x))+\frac{1}{\alpha}\phi_{\alpha x}(s,\zeta_\varepsilon(s,t,x))\right]\right\rangle ds.\end{aligned}\tag{4.15}$$

By assumption (iii) the first two terms of (4.15) are positive. Concerning the last two terms we show that they are smaller than the second one. We have by the inequalities (4.3)-(4.6):

$$\begin{aligned}&|e^{-t/\alpha}\langle \zeta_{\varepsilon xx}(0,t,x)(z,z), \phi_0'(\zeta_\varepsilon(0,t,x))\rangle| \\ &\qquad \leqq e^{-t/\alpha}|\phi_0|_{1,R}|f(\varepsilon,\cdot)|_{2,R}\int_0^t |\zeta_{\varepsilon x}(s,t,x)z|^2\,ds \overset{\text{def}}{=} A_{\varepsilon,R}\end{aligned}$$

and

$$\begin{aligned}&\left|\int_0^t e^{-(t-s)/\alpha}\left\langle \zeta_{\varepsilon xx}(s,t,x)(z,z), \left[g'(\zeta_\varepsilon(s,t,x))+\frac{1}{\alpha}\,\phi_{\alpha x}(s,t,x))\right]\right\rangle ds\right| \\ &\qquad \leqq e^{-t/\alpha}\int_0^t e^{s/\alpha}\varepsilon\left(|g|_{1,R}+\frac{1}{\alpha}\,C\right)|f(\varepsilon,\cdot)|_{2,R}\left(\int_s^t |\zeta_{\varepsilon x}(\sigma,t,x)z|^2\,dx\right)ds \\ &\qquad \leqq e^{-t/\alpha}\left(|g|_{1,R}+\frac{1}{\alpha}\,C\right)|f(\varepsilon,\cdot)|_{2,R}\int_0^t |\zeta_{\varepsilon x}(\sigma,t,x)z|^2\left(\int_0^\sigma e^{s/\alpha}\,ds\right)d\sigma \\ &\qquad \leqq e^{-t/\alpha}(\alpha|g|_{1,R}+C)|f(\varepsilon,\cdot)|_{2,R}\int_0^t e^{s/\alpha}|\zeta_{\varepsilon x}(s,t,x)z|^2\,ds \overset{\text{def}}{=} B_{\varepsilon,R}.\end{aligned}$$

It follows that

$$A_{\varepsilon,R}+B_{\varepsilon,R}\leqq |f(\varepsilon,\cdot)|_{2,R}(|\phi_0|_{1,R}+|g|_{1,R}+C)\int_0^t e^{-(t-s)/\alpha}|\zeta_{\varepsilon x}(s,t,x)z|^2\,ds.$$

Set $\varepsilon(R,C)$ such that for all $|\varepsilon|\leqq\varepsilon(R,C)$, we have

$$|f(\varepsilon,\cdot)|_{2,R}\leqq\frac{\mu}{|\phi_0|_{1,R}+|g|_{1,R}+C}, \tag{4.16}$$

then by assumption (iii) it follows:

$$\langle(\Gamma_\alpha^\varepsilon\phi)_{xx}(t,x)z,z\rangle\geqq 0.$$

Finally, by (4.13) we immediately get

$$(\Gamma_\alpha^\varepsilon\phi)_x(t,0)=0 \quad \forall t\in[0,T]$$

and the proof is complete. □

4.4. Successive approximations.

LEMMA 4.3. *Let* (i)-(iii) *hold true. Fix* $R>0$, *and set*

$$L=L(R)=|\phi_0|_{C^2_{\text{Lip}}(\Sigma_r)}+T|g|_{C^2_{\text{Lip}}(\Sigma_R)}.$$

Let $|\varepsilon|\leqq\varepsilon(R,L)$. *Consider the sequence* $\{\phi^n\}_{n\in\mathbf{N}}\subset B([0,T];C^2_{\text{Lip}}(\Sigma_R)\cap K_R)$ *defined by*

$$\phi^0=\Gamma_\alpha^\varepsilon(0), \qquad \phi^n=\Gamma_\alpha^\varepsilon(\phi^{n-1}),$$

Then there exists $\varepsilon_2(R)\in(0,\varepsilon(R,L))$ *such that for* $|\varepsilon|\leqq\varepsilon_2(R)$, *we have*:

(I) $\phi\in\mathscr{F}_C^R$ *and also* $\sup_{t\in[0,T]}|\phi^n(t,\cdot)|_{C^2_{\text{Lip}}(\Sigma_R)}\leqq L(R)$, *for all* $n\in\mathbf{N}$;
(II) $|\phi^n(t,\cdot)-\phi^{n-1}(t,\cdot)|_{C^2(\Sigma_R)}\leqq(L(R)+1)^{n+1}t^n/\alpha^n n!$, *for all* $n\in\mathbf{N}$;
(III) $\Gamma_\alpha^\varepsilon$ *is continuous on* $B([0,T];C^2(\Sigma_R))$.

Remark 4.4. This implies that there exists $\phi_\varepsilon^\alpha\in B([0,T];C^2_{\text{Lip}}(\Sigma_R)\cap K_R)$ such that

$$\phi^n\xrightarrow{n\to\infty}\phi_\varepsilon^\alpha \quad \text{in } B([0,T];C^2(\Sigma_R))$$

and therefore (by (III))

$$\phi^\alpha_\varepsilon(t,x)=(\Gamma^\varepsilon_\alpha\phi^\alpha_\varepsilon)(t,x)\quad \text{on } [0,T]\times\Sigma_R$$

and ϕ^α_ε is also a strict solution of $(\mathrm{HJ})_\alpha$.

We also remark that estimate (I) does not depend on α. This fact will be used to prove the convergence of ϕ^α as $\alpha\to 0$ in the next section.

Proof of (I). We use a recurrence argument. First we have that (I) holds true for the function $0\in B([0,T]; C^2_{\mathrm{Lip}}(\Sigma_R))$. After we suppose that (I) holds true for $\phi^0,\cdots,\phi^{n-1}$. Then by using (4.12)-(4.14), and Lemma 2.1, we obtain

$$|\phi^n(t,\cdot)|_{C^2_{\mathrm{Lip}}(\Sigma_R)}\leqq e^{-t/\alpha}|\phi_0|_{C^2_{\mathrm{Lip}}(\Sigma_R)}\gamma(0,t)+|g|_{C^2_{\mathrm{Lip}}(\Sigma_R)}\int_0^t e^{-(t-s)/\alpha}\gamma(s,t)\,ds$$
$$+\frac{1}{\alpha}\int_0^t e^{-(t-s)/\alpha}|\phi^{n-1}(s,\cdot)|_{C^2_{\mathrm{Lip}}(\Sigma_R)}\gamma(s,t)\,ds,$$

where

$$\gamma(s,t)\leqq e^{+\omega(t-s)/2}+(|f(\varepsilon,\cdot)|_{2,R}+|f(\varepsilon,\cdot)|^2_{2,R}T+\|f(\varepsilon,\cdot)\|_{2,R})(t-s).$$

It is easy to see that, if we set $E_R=|f(\varepsilon,\cdot)|_{2,R}+|f(\varepsilon,\cdot)|^2_{2,R}T+\|f(\varepsilon,\cdot)\|_{2,R}$, then for ε sufficiently small we have $E_R\leqq(-\omega/2)\,e^{T\omega/2}$ and hence $\gamma(s,t)\leqq 1$, for all s, $t\in[0,T]$. Now by iterating the above estimate we obtain

$$|\phi^n(t,\cdot)|_{C^2_{\mathrm{Lip}}(\Sigma_R)}\leqq e^{-t/\alpha}|\phi_0|_{C^2_{\mathrm{Lip}}(\Sigma_R)}\sum_{k=0}^n\frac{1}{k!}\frac{t^k}{\alpha^k}$$
$$+|g|_{C^2_{\mathrm{Lip}}(\Sigma_R)}\int_0^t e^{-(t-s)/\alpha}\sum_{k=0}^n\frac{1}{k!}\frac{(t-s)^k}{\alpha^k}\,ds\leqq L(R).$$

Proof of (II). Using the same method as (I) we get

$$|\phi^n(t,\cdot)-\phi^{n-1}(t,\cdot)|_{C^2(\Sigma_R)}\leqq\frac{1}{\alpha}\int_0^t e^{-(t-s)/\alpha}\gamma_1(s,t)|\phi^{n-1}(s,\cdot)-\phi^{n-2}(s,\cdot)|_{C^2(\Sigma_R)}\,ds$$

and it is easy to prove that

$$\gamma_1(s,t)\leqq L(R)+1.$$

By iterating we get

$$|\phi^n(t,\cdot)-\phi^{n-1}(t,\cdot)|_{C^2(\Sigma_R)}\leqq\frac{L(R)}{\alpha}\int_0^T e^{-(t-s)/\alpha}\frac{(L(R)+1)^n}{(n-1)!}\frac{(t-s)^{n-1}}{\alpha^{n-1}}\,ds$$

from which (II) follows.

Proof of (III). Let ϕ, $\bar\phi\in B([0,T]; C^2_{\mathrm{Lip}}(\Sigma_R)\cap K_R)$, and let

$$D=\sup_{t\in[0,T]}\{|\phi(t,\cdot)|_{C^2(\Sigma_R)},|\bar\phi(t,\cdot)|_{C^2(\Sigma_R)}\}.$$

Then we have

$$\sup_{t\in[0,T]}\{|\Gamma^\varepsilon_\alpha\phi(t,\cdot)-\Gamma^\varepsilon_\alpha\bar\phi(t,\cdot)|_{C^2(\Sigma_R)}\}$$
$$\leqq\frac{1}{\alpha}\int_0^t e^{-(t-s)/\alpha}(D+1)|\phi(s,\cdot)-\bar\phi(s,\cdot)|_{C^2(\Sigma_R)}\,ds$$
$$\leqq\frac{(D+1)T}{\alpha}\sup_{t\in[0,T]}|\phi(t,\cdot)-\bar\phi(t,\cdot)|_{C^2(\Sigma_R)}$$

and (III) is proved. □

Remark 4.5. If $f(\varepsilon,\cdot)\in C^k_{\mathrm{Lip}}(H,H)$ and g, $\phi_0\in C^k(H)$ $(k\geqq 3)$ it is possible to prove (by reducing the limit for $|\varepsilon|$) that $\phi^n\in B([0,T];\ C^k_{\mathrm{Lip}}(\Sigma_R)\cap K_R)$ and

$$\phi^n \xrightarrow{n-\infty} \phi^\alpha_\varepsilon \quad \text{in } B[0,T];\ C^k(\Sigma_R)).$$

4.5. Convergence of ϕ^α_ε as $\alpha\to 0$.

LEMMA 4.6. *We assume that* (i)–(iii) *hold. Let $R>0$ be fixed and let $\varepsilon_2(R)$ be as in* § 4.4. *Let $\phi^\alpha_\varepsilon\in B([0,T];\ C^2(\Sigma_R)\cap K_R)$ be the strict solution of* $(\mathrm{HJ})_\alpha$.

Then there exists $\phi_\varepsilon\in B([0,T];\ C^1_{\mathrm{Lip}}(\Sigma_R)\cap K_R)$ such that:

(a) $\phi^\alpha_\varepsilon \xrightarrow{\alpha\to 0} \phi_\varepsilon$ *in* $B([0,T];\ C^1(\Sigma_R))$,

(b) ϕ_ε *is a strict solution of the Hamilton-Jacobi equation* (HJ),

(c) *for all* $t\in[0,T]$, $\phi_\varepsilon(t,\cdot)\in C^2_{\mathrm{Lip}}(\Sigma_R)\cap K_R)$.

Proof. Due to the estimates in § 4.4 (see Remark 4.4) the proof is completely analogous to the one of [2, pp. 38–41, 98–99]; thus it will be only sketched.

Let $\alpha>0$, $\beta>0$. Then ϕ^β_ε (we write ϕ^β for simplicity) fulfills the equation

$$\phi^\beta_t(t,x)+\frac{1}{\alpha}(\phi^\beta-\phi^\beta_\alpha)-\langle Ax+f(\varepsilon,x),\ \phi^\beta_x\rangle=g(x)+R_{\alpha,\phi^\beta}+R_{\beta,\phi^\beta} \tag{4.17}$$

$$\forall(t,x)\in[0,T]\times D(A)\cap\Sigma_R,$$

$$\phi^\beta(0,x)=\phi_0(x)\quad \forall x\in H.$$

It follows, by using the integral form, that

$$|\phi^\beta(t,\cdot)-\phi^\alpha(t,\cdot)|_{C^1(\Sigma_R)}\leqq\frac{1}{\alpha}\int_0^t e^{-(t-s)/\alpha}|\phi^\beta(s,\cdot)-\phi^\alpha(s,\cdot)|_{C^1(\Sigma_R)}\,ds$$

$$+\int_0^t e^{-(t-s)/\alpha}(|R_{\alpha,\phi^\beta(s,\cdot)}|_{C^1(\Sigma_R)}+|R_{\beta,\phi^\beta(s,\cdot)}|_{C^1(\Sigma_R)}\,ds.$$

Now by Lemma 2.2 and Gronwall's inequality we get

$$|\phi^\beta(t,\cdot)-\phi^\alpha(t,\cdot)|_{C^1(\Sigma_R)}\leqq(\alpha+\beta)3L^3(R)T,$$

so (a) is proved.

To show (b) we remark that, if $x\in D(A)$, then $\phi^\alpha(\cdot,x)\in C^1([0,T])$, and

$$\phi^\alpha_t(t,x)=g(x)+R_{\alpha,\phi^\alpha(t,\cdot)}(x)-\tfrac{1}{2}|\phi^\alpha_x(t,x)|^2+\langle Ax+f(\varepsilon,x),\ \phi^\alpha_x(t,x)\rangle,$$

so $\phi^\alpha_t(\cdot,x)\xrightarrow{\alpha\to 0}\phi_{\varepsilon t}$ in $C([0,T])$, and ϕ_ε is a strict solution.

To show (c) we apply Ascoli's theorem exactly as in [2, pp. 40–41]. □

Remark 4.7. As in the previous section, if $f(\varepsilon,\cdot)\in C^k_{\mathrm{Lip}}(H,H)$, g, $\phi_0\in C^k(H)$ and $|\varepsilon|$ is sufficiently small, then it is possible to prove that

$$\phi^\alpha_\varepsilon\xrightarrow{\alpha\to 0}\phi_\varepsilon\quad\text{in } B([0,T];\ C^{k-1}(\Sigma_R))$$

and

$$\phi_\varepsilon(t,\cdot)\in C^k_{\mathrm{Lip}}(\Sigma_R)\cap K_R\quad\forall t\in[0,T].$$

4.6. Uniqueness. This also is standard and we only sketch the proof (see [2, pp. 39, 99], [4] for more details). Let ϕ_1, ϕ_2 be two strict solutions of (3.1). Then for $\alpha>0$ we have

$$\phi_{i,t}(t,x)=-\frac{1}{\alpha}(\phi_i-\phi_{i,\alpha})+\langle Ax+f(\varepsilon,x),\ \phi_{i,x}\rangle+g(x)+R_{\alpha,\phi_i}$$

$$\forall(t,x)\in[0,T]\times D(A)\cap\Sigma_R,\quad i=1,2,$$

$$\phi_i(0,x)=\phi_0(x)\quad\forall x\in H.$$

By using the integral form (4.1) and applying Gronwall's inequality we obtain

$$|\phi_1(t,\cdot)-\phi_2(t,\cdot)|_{C^1(\Sigma_R)}\leqq 5\alpha L^3(R)T$$

and the uniqueness follows from the arbitrariness of α.

Now Theorem 3.2 is proved and we may set $\bar{\varepsilon}(R)=\varepsilon_2(R)$ as in § 4.4.

Remark 4.8. At this point, as we said in the beginning of this section, we show that the result of Theorem 3.2 also holds true when the operator A is not strictly dissipative. In this case we have, by the hypothesis (i):

$$\langle Ax, x\rangle \leqq \omega|x|^2 \quad \forall x\in H \quad \text{for some } \omega\geqq 0.$$

We proceed in this way (see [2, Remark 5, p. 46]. Set $\eta(t,x)=\phi(t, e^{-2\omega t}x)$. Then (HJ) reduces to

$$(\overline{\text{HJ}})\qquad \begin{aligned}&\eta_t(t,x)=-\tfrac{1}{2}b(t)|\eta_x(t,x)|^2+\langle A_1x+f_1(\varepsilon,x,t),\eta_x(t,x)\rangle+g_1(t,x)\\ &\qquad\qquad\qquad\qquad\qquad\qquad \forall (t,x)\in[0,T]\times D(A),\\ &\eta(0,x)=\phi_0(x)\quad \forall x\in H,\end{aligned}$$

where we posed

$$b(t)=e^{4\omega t},\qquad A_1=A-2\omega,$$

$$f_1(\varepsilon,x,t)=e^{2\omega t}f(\varepsilon, x\,e^{-2\omega t}),\qquad g_1(t,x)=g(x\,e^{-2\omega t}).$$

Now we can write $(\overline{\text{HJ}})$ in approximating integral form as follows:

$$(4.18)\qquad \begin{aligned}\eta(t,x)=e^{-B(t)}\phi_0(\zeta_\varepsilon^1(0,t,x))&+\int_0^t e^{-(B(t)-B(s))}\\ &\cdot\left[g_1(s,\zeta_\varepsilon^1(s,t,x))+\frac{1}{\alpha}b(s)\eta_\alpha(s,\zeta_\varepsilon^1(s,t,x))\right]ds,\end{aligned}$$

where

$$B(t)=\frac{1}{\alpha}\int_0^t b(s)\,ds$$

and $\zeta_\varepsilon^1(s,t,x)$ is the solution of the Cauchy problem,

$$(4.19)\qquad \begin{aligned}&(\zeta_\varepsilon^1)'(s)=-A_1\zeta_\varepsilon^1(s)-f_1(\varepsilon,\zeta_\varepsilon^1(s),s)\quad \text{for } s\in[0,T],\\ &\zeta_\varepsilon^1(s)=x,\qquad x\in H.\end{aligned}$$

Now we can solve this approximating problem using the same method seen in §§ 4.2, 4.3, and 4.4. The only differences are:

(1) The functions g_1, f_1 are time-dependent, but this dependence is uniform on $[0,T]\times\Sigma_R$ for all $R>0$ (also for the derivatives), so it is sufficient to change the value of the constants which appear in the proof. In this case all these constants depend on T and they blow up when T goes to $+\infty$.

(2) The factor $1/\alpha$ is replaced by $b(t)/\alpha$ and t/α by $B(t)=(1/\alpha)\int_0^t b(s)\,ds$. However, we can repeat all of our estimates using that $b(t)$ is bounded on $[0,T]$ and $b(t)b(s)=b(t+s)$.

Finally, the proofs of §§ 4.5 and 4.6 are completely analogous by arranging the constants.

5. Solution of the control problem. In this section we apply the results of Theorem 3.2 to the solution of the control problem (P). We assume (i)-(iii) and for simplicity we discuss only the case in which the operator A is strictly dissipative (that is, $\omega < 0$ in the hypothesis (i)). Anyway, all the results of this section also hold true in the general case. The proof for the general case is the same: we have only to arrange the constants. Let r, $M > 0$ be fixed. Suppose $|x| \leqq r$, $|u|_{L^2(0,T;H)} \leqq L(r)$. Now set $R = r + \sqrt{T}M$ and $|\varepsilon| \leqq \bar{\varepsilon}$.

Then, as seen in § 3, the following statements hold true:

(1) The state equation (1) has a unique mild solution

$$y \in C([0, T]; H) \quad \text{with } |y(t)| \leqq R.$$

(2) The Hamilton-Jacobi equation (HJ) has a unique strict solution

$$\phi_\varepsilon \in B([0, T]; C^1_{\mathrm{Lip}}(\Sigma_R) \cap K_R).$$

Using these results we can solve the control problem (P) in a standard way.

LEMMA 5.1. *Let r, $M > 0$ be fixed. Let $x \in \Sigma_r$, $u \in L^2(0, T; H)$, with $|u|_{L^2} \leqq M$. Let $R = r + \sqrt{T}M$ and $|\varepsilon| \leqq \bar{\varepsilon}(R)$. Let y be the mild solution of the Cauchy problem*:

$$(5.1) \qquad \begin{aligned} &y' = Ay + f(\varepsilon, y) + u \quad \text{on } [0, T], \\ &y(t) = x. \end{aligned}$$

Then the following fundamental identity holds for every $(t, x) \in [0, T] \times \Sigma_r$:

$$(5.2) \qquad \begin{aligned} &\phi_\varepsilon(T - t, x) + \frac{1}{2}\int_t^T |u + \phi_{\varepsilon x}(T - s, y(s))|^2 \, ds \\ &= \phi_0(y(T)) + \int_t^T \left[g(y(s)) + \frac{1}{2}|u(s)|^2 \right] ds \\ &= J(x, u) \quad \text{if } t = 0. \end{aligned}$$

The proof is standard (see, for instance, [2, pp. 51-52], [7]). We only recall that for $|\varepsilon| < \bar{\varepsilon}(r + \sqrt{T}M)$, expression (5.2) makes sense, due to statements (1) and (2).

Now we consider the closed-loop equation:

$$(5.3) \qquad \begin{aligned} &y'(s) = Ay(s) + f(\varepsilon, y(s)) - \phi_{\varepsilon x}(T - s, y(s)) \quad \text{on } [0, T], \\ &y(t) = x, \qquad x \in \Sigma_r. \end{aligned}$$

We remark that, since $\phi_{\varepsilon x}(T - t, \cdot)$ is a locally Lipschitz monotone operator on H, then (5.3) has a unique mild solution $y_\varepsilon \in C([t, T]; H)$. Furthermore, due to the monotonicity of $\phi_{\varepsilon x}(T - t, \cdot)$ we have that

$$(5.4) \qquad |y_\varepsilon(s)| \leqq |x| \leqq r \quad \forall s \in [t, T].$$

Hence we can state the following theorem.

THEOREM 5.2. *Let $r_0 > 0$ be fixed. There exists $\varepsilon_3(r_0) > 0$ such that for $|x| \leqq r_0$, $|\varepsilon| \leqq \varepsilon_3(r_0)$, problem* (P) *has a unique optimal control u_ε. Moreover, u_ε is given by the feedback formula*

$$(5.5) \qquad u_\varepsilon(s) = -\phi_{\varepsilon x}(T - s, y_\varepsilon(s)),$$

where $y_\varepsilon(s)$ is the solution of the closed-loop equation (5.3) *for $t = 0$.*

Proof. Let

$$M_0 = M_0(r_0) = \max \{\sqrt{T}L(r_0), L(r_0), \sqrt{2L(r_0)}\}$$

as in Lemma 4.3. First we observe that if $|u|_{L^2} > M_0$, then $J(u) > J(0)$, in fact,

$$J(x, 0) = \int_0^T g(y_1(s))\,ds + \phi_0(y_1)(T)),$$

where

$$y_1' = Ay_1 + f(\varepsilon, y_1). \qquad y_1(0) = x,$$

which implies

$$|y_1(s)| \leqq |x|\, e^{+\omega s/2} \quad \forall |\varepsilon| \leqq \varepsilon_1(r_0)$$

and

$$J(x, 0) \leqq |g|_{0,r_0} T + |\phi_0|_{0,r_0} \leqq L(r_0).$$

But

$$J(x, u) \geqq |u|_{L^2}^2 - |\phi_0(0)| - T|g(0)| \geqq M_0^2 - L(r_0)$$
$$\geqq J(x, 0) \quad \text{as } |u|_{L^2(0,T;H)} > M_0.$$

So we only have to minimize J for $u \in L^2(0, T; H)$, $|u|_{L^2} \leqq M_0$.

Now set $\varepsilon_3(r_0) = \bar{\varepsilon}(r_0 + \sqrt{T}\, M_0)$ and let $|\varepsilon| \leqq \varepsilon_3(r_0)$. At this point we need only to show that u_ε given by (5.5) is such that $|u_\varepsilon|_{L^2} \leqq M_0(r_0)$. In fact, in this case it easily follows by Lemma 5.1 that u_ε is the unique optimal control for (P) (see [2], pp. 50–54].

We have by (5.4)

$$|\phi_{\varepsilon x}(T - s, y_\varepsilon(s))| \leqq \sup_{t \in [0,T]} |\phi_\varepsilon(T - s, \cdot)|_{1,r_0} \leqq L(r_0).$$

It follows that

$$|u_\varepsilon|_{L^2} \leqq \sqrt{T} L(r_0) \leqq M_0(r_0). \qquad \square$$

6. Regularity with respect to the parameter ε. In this section we give a regularity result for the solution of our control problem.

First of all we remark that if $f \in C^k_{\text{Lip}}(H, H)$, then it is not difficult to prove that the solution ϕ_ε of the Hamilton–Jacobi equation (HJ) and its spatial derivative $\phi_{\varepsilon x}$ are $(k-1)$-times differentiable with respect to ε (see Remarks 4.5 and 4.7).

For simplicity we assume that, for every $x \in H$,

$$\text{(iv)} \quad \begin{aligned} &\phi_0(x) = \tfrac{1}{2}\langle P_0 x, x\rangle, \\ &g(x) = \tfrac{1}{2}\langle Mx, x\rangle, \\ &f(\varepsilon, x) = \varepsilon f(x), \\ &\omega < 0 \quad \text{in hypothesis (i)} \end{aligned}$$

with P_0, $M \in \Sigma^+(H)$, $f \in C^2_{\text{Lip}}(H, H)$, and $\langle Mx, x\rangle \geqq \mu |x|^2$, for all $x \in H$. Then if $\varepsilon = 0$, the solution of Hamilton–Jacobi equation (HJ) is the quadratic form $x \to \frac{1}{2}\langle P(t)x, x\rangle$, where $P(t)$ is the solution of the Riccati operator equation

$$\text{(6.1)} \quad \begin{aligned} &P' = A^*P + PA - P^2 + M \quad \text{on } [0, T], \\ &P(0) = P_0. \end{aligned}$$

PROPOSITION 6.1. *Let* (i)-(iv) *hold. Fix any* $R>0$, *and let* $|\varepsilon|\leqq\bar{\varepsilon}(R)$. *Then, for every* $n\in\mathbf{N}$ *and for every* $x\in\Sigma_R$ *there exist the derivatives*:

$$\frac{\partial^n}{\partial\varepsilon^n}\phi_\varepsilon(t,x)|_{\varepsilon=0},\qquad \frac{\partial^n}{\partial\varepsilon^n}\phi_{\varepsilon x}(t,x)|_{\varepsilon=0},$$

and we have

$$\frac{1}{n!}\frac{\partial^n}{\partial\varepsilon^n}\phi_\varepsilon(t,x)|_{\varepsilon=0}=\phi_n(t,x),$$

$$\frac{1}{n!}\frac{\partial^n}{\partial\varepsilon^n}\phi_{\varepsilon x}(t,x)|_{\varepsilon=0}=\phi_{n,x}(t,x),$$

where $\phi_0(t,x)=\frac{1}{2}\langle P(t)x,x\rangle$, *and for* $n\geqq1$, ϕ_n *is the strict solution of the equation*

$$\begin{aligned}&\phi_{nt}=\langle(A-P(t))x,\phi_{n,x}\rangle+\langle f(x),\phi_{(n-1),x}\rangle-\sum_{h=1}^{n-1}\langle\phi_{h,x},\phi_{(n-h),x}\rangle,\\&\phi_n(0,x)=0.\end{aligned}\tag{6.2}$$

Moreover, the following recursive formula holds:

$$\begin{aligned}\phi_n(t,x)=&\int_0^t\langle f(U(t-s,0)x),\phi_{(n-1),x}(s,U(t-s,0)x)\rangle\,ds\\&-\int_0^t\sum_{h=1}^{n-1}\langle\phi_{h,x}(s,U(t-s,0)x),\phi_{(n-h),x}(s,U(t-s,0)x)\rangle\,ds,\end{aligned}\tag{6.3}$$

where $U(t,s)$ *is the evolution operator associated to* $A-P(t)$. *Similar formulas hold true for the derivatives with respect to* ε *of the function* $\phi_{\varepsilon x}$ (*obtained by differentiating with respect to* x *the corresponding formulas for* ϕ_ε).

Proof. For every $n\in\mathbf{N}$ we consider the function

$$\delta_n=\frac{1}{\varepsilon^n}\left(\phi_\varepsilon-\frac{1}{2}\langle P(t)x,x\rangle-\varepsilon\phi_1-\cdots-\varepsilon^n\phi_n\right)$$

and we show that recursively

$$\left.\begin{aligned}&\lim_{\varepsilon\to0}\delta_n(t,x)=0\\&\lim_{\varepsilon\to0}\delta_{n,x}(t,x)=0\end{aligned}\right\}\text{uniformly on }[0,T]\times\Sigma_R.\tag{6.4}$$

For instance, in the case where $n=0$, we have that the function $\delta_0(t,x)=\phi_\varepsilon(t,x)-\frac{1}{2}\langle P(t)x,x\rangle$ satisfies the following Cauchy problem in H:

$$\begin{aligned}&\delta_{0t}(t,x)=\langle F(t,x),\delta_{0x}(t,x)\rangle+\varepsilon G(t,x)\quad\forall(t,x)\in[0,T]\times(D(A)\cap\Sigma_R),\\&\delta(0,x)=0\quad\forall x\in\Sigma_R,\end{aligned}\tag{6.5}$$

where

$$F(t,x)=Ax-\tfrac{1}{2}[P(t)x-\phi_{\varepsilon x}(t,x)]$$

and

$$G(t,x)=\langle f(x),P(t)x\rangle.$$

Using the classical argument of characteristics, we obtain that

$$\delta_0(t,x)=\varepsilon\int_0^t G(s,\xi(s,t,x))\,ds\quad\forall(t,x)\in[0,T]\times\Sigma_R,\tag{6.6}$$

where $\xi(s, t, x)$ is the solution of the Cauchy problem in H:

$$\begin{aligned}&\frac{\partial \xi}{\partial s}(s, t, x) = -F(s, \xi(s, t, x)) \quad \text{for } s \in [0, t],\\ &\xi(t, t, x) = x \quad \text{for } x \in \Sigma_R.\end{aligned} \tag{6.7}$$

We observe that, due to the hypothesis (iv) and the monotonicity of $\phi_{\varepsilon x}$ the Cauchy problem (6.7) has a unique mild solution $\xi(s, t, x) \in C([0, T]; H)$ for every $x \in \Sigma_R$, and that

$$|\xi(s, t, x)| \leqq |x|.$$

This enables us to write (6.6) for every $x \in \Sigma_R$. Now by (6.6) and boundedness of G on $[0, T] \times \Sigma_R$ it follows that

$$\lim_{\varepsilon \to 0} \delta_0(t, x) = 0 \quad \text{uniformly on } [0, T] \times \Sigma_R.$$

A similar argument can be repeated for the spatial derivative of δ_0 using that the function $E(t, x) = \phi_{\varepsilon x}(t, x)$ satisfies the quasi-linear operator equation (see [2, p. 100]

$$\begin{aligned}&E_t(t, x) = [A + \varepsilon f'(x)]^* E(t, x) + E_x(t, x)[Ax + \varepsilon f(x) - E(t, x)] + M\\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{on } [0, T] \times (D(A) \cap \Sigma_R)\\ &E(0, x) = P_0.\end{aligned} \tag{6.8}$$

Finally, we can recursively repeat the same argument also for every $n \in \mathbf{N}$. □

REMARK 6.2. Proposition 6.1 also holds if we have

$$\phi_0(x) = \tfrac{1}{2}\langle P_0 x, x\rangle + \theta_\varepsilon(x), \qquad g(x) = \tfrac{1}{2}\langle Mx, x\rangle + \eta_\varepsilon(x),$$

where θ, η are smooth function with respect to ε, vanishing for $\varepsilon = 0$. Moreover, θ, η must belong to $C^2_{\text{Lip}}(H, H)$ with respect to x.

Note also that we have not used any additional regularity hypothesis on f.

Finally, we have Proposition 6.3.

PROPOSITION 6.3. *Let* (i)–(iv) *hold and let* $f \in C^k_{\text{Lip}}(H, H)$ $(k \geqq 2)$. *Then for* $\varepsilon \in [-\bar{\varepsilon}(R), \bar{\varepsilon}(R)]$, *and* $h = 1, \cdots, k-1$ *there exist* $(\partial^h/\partial \varepsilon^h)u_\varepsilon$, $(\partial^h/\partial \varepsilon^h)y_\varepsilon$ *and*

$$\frac{1}{h!}\frac{\partial^h}{\partial \varepsilon^h} u_\varepsilon(s) = u_h(s), \qquad \frac{1}{h!} = \frac{\partial^h}{\partial \varepsilon^h} y_\varepsilon(s) = y_h(s),$$

where we have

$$y_0'(s) = (A - P(T-s))y_0(s), \qquad y_0(0) = x,$$

$$y_1'(s) = (A - P(T-s))y_1(s) + f(y_0(s)) - \psi_{1,x}(s, y_0(s)), \qquad y_1(0) = 0,$$

and, for $h > 1$,

$$\begin{aligned} y_h'(s) = (A - P(T-s))y_h(s) &- \phi_{h,x}(s, y_0(s))\\ &+ \sum_{r=1}^{h-1} \sum_{\substack{i_1 + \cdots + i_r = h-1\\ i_j \geqq 1}} \frac{1}{r!}\frac{\partial^r}{\partial x^r} f(y_0(s))(y_{i_1}(s), \cdots, y_{i_r}(s))\\ &+ \sum_{q=1}^{h} \sum_{r=1}^{h-q} \sum_{\substack{i_1 + \cdots + i_r = h-q\\ i_j \geqq 1}} \frac{1}{r!}\frac{\partial^r}{\partial x^r} \psi_{q,x}(s, y_0(s))(y_{i_1}(s), \cdots, y_{i_r}(s))\end{aligned}$$

$$y_h(0) = 0.$$

Moreover,

$$u_0(s) = -P(T-s)y_0(s),$$

and, for $h>0$,

$$u_h(s) = -\phi_{h,x}(s, y_0(s)) - \sum_{q=1}^{h} \sum_{r=1}^{h-q} \sum_{\substack{i_1+\cdots+i_r=h-q \\ i_j \geqq 1}} \frac{1}{r!} \frac{\partial^r}{\partial x^r} \psi_{q,x}(s, y_0(s))(y_{i_1}(s), \cdots, y_{i_r}(s))$$

For the proof we argue as in Proposition 6.1.

7. Examples. We give two examples.

Parabolic systems. Let $H = H^1(0, 1)$. We consider the parabolic state equation:

$$\text{(7.1)} \quad \begin{aligned} &y_t(t, x) = \Delta_x y(t, x) + \varepsilon f(y(t, x)) + u(t, x), \qquad t \in [0, T], \quad x \in (0, 1), \\ &y(0, x) = y_0(x) \in H^1(0, 1), \quad y(t, 0) = 0 = y(t, 1) \quad \forall t \in [0, T]. \end{aligned}$$

We denote by A the operator on H defined by

$$D(A) = H^3(0, 1) \cap H_0^1(0, 1), \qquad Ay = \Delta y.$$

The control u is any element of $L^2(0, T; H)$, and f is a smooth function $\mathbf{R} \to \mathbf{R}$ such that, if we define

$$F: H \to H, \quad F(y)(x) = f(y(x)) \quad \forall x \in (0, 1),$$

then we have $f \in C^2_{\text{Lip}}(H, H)$. For example, we can take $f(z) = z^2$ and use the fact that $H^1(0, 1)$ is an algebra.

We want to minimize the cost

$$\text{(72)} \qquad J(y_0, u) = \int_0^T \left[\frac{1}{2}|y(s)|_H^2 + \frac{1}{2}|u(s)|_H^2\right] ds + \frac{1}{2}|y(T)|_H^2$$

over all controls $u \in L^2(0, T; H)$, where y is the mild solution of (7.1).

Now assumptions (i)-(iv) are verified. In particular,

$$\langle Ay, y\rangle_H \leqq -C_0|y|_H^2 \quad \forall y \in D(A),$$

where $C_0 > 0$.

If r_0 is the supremum of the norm of the initial state y_0, we take M_0 as in § 5, and we set $|\varepsilon| \leqq \varepsilon_3(r_0)$.

Then by the Theorems 3.2 and 5.2 there exists a unique optimal pair $(u_\varepsilon, y_\varepsilon)$ for problem (P), and the following feedback formula holds:

$$u_\varepsilon(s) = -\phi_{\varepsilon x}(T-s, y_\varepsilon(s)),$$

where ϕ_ε is the strict solution of the Hamilton-Jacobi equation (HJ).

Furthermore, if $F \in C^k_{\text{Lip}}(H, H)$ (which is true if $f(z) = z^2$), then the optimal pair is k-times differentiable with respect to ε and the expansion given in Proposition 6.3 holds.

We can repeat the same example in higher dimensions by setting $H = H^k(\Omega)$, where Ω is an open bounded subset of $\mathbf{R}^n$ with sufficiently smooth boundary $\partial\Omega$ and $k > n/2$ (so H is an algebra).

Hyperbolic systems. Consider the following optimization problem:

Minimize

$$J(y_0, y_1, u) = \int_0^T \left[g(y(s), y'(s)) + \frac{1}{2}|u(s)|^2\right] ds + \phi_0(y(T), y'(T))$$

subject to

$$y''(s)+A_0(y)(s)+\varepsilon f(y(s))=u(s), \qquad s\in[0,T],$$
$$y(0)=y_0, \quad y'(0)=y_1, \quad y_0\in V, \quad y_1\in E,$$

where V, E are real Hilbert spaces such that $V\subset E\subset V'$ and the inclusion $V\to E$ is continuous and densely defined (V' is the dual of V and E is identified with its own dual). Moreover, $u\in L^2(0,T;E)$.

$A_0\colon V\to V'$ is a linear continuous symmetric operator, and denoting by $\langle\cdot,\cdot\rangle_0$ the duality between V and V', there exists $\omega>0$ such that

$$\langle A_0y,y\rangle_0\geqq\omega|y|^2 \quad \forall y\in V.$$

$f\colon V\to E$ is a twice Fréchet differentiable mapping with $f^{(k)}$ $(k=0,1,2)$ locally bounded and locally Lipschitz continuous. Finally, concerning the functions g and ϕ_0 we assume that

$$g,\phi_0\in C^2_{\mathrm{Lip}}(V\times E), \quad \phi \text{ is convex}, \quad g \text{ is strictly convex}.$$

We set $H=V\times E$ and we endow H of the usual Hilbert structure:

$$\langle(v_1,e_1),(v_2,e_2)\rangle=\langle A_0v_1,v_2\rangle+\langle e_1,e_2\rangle \quad \text{for } v_1,v_2\in V, \quad e_1,e_2\in E.$$

Setting $Y(s)=(y(s),y'(s))$ and $Y_0=(y_0,y_1)$, we can write our state equation as

$$Y'(s)+AY(s)+\varepsilon F(Y(s))=U(s) \quad \text{for } s\in[0,T],$$
$$Y(0)=Y_0, \qquad Y_0\in H,$$

where

$$A=\begin{pmatrix}0 & -1\\ A_0 & 0\end{pmatrix}, \quad U=(0,u), \quad F(y,z)=(0,f(y)) \quad \forall(y,z)\in V\times E=H.$$

Now, our problem is expressed in the form (1.2) in terms of the operators A and F, and (i)-(iii) hold. We can proceed as in the previous example to solve the control problem.

We recall a typical example of this situation. Let $E=L^2(\Omega)$, where Ω is an open bounded subset of $\mathbf{R}^3$ with sufficiently smooth boundary $\partial\Omega$. Moreover, let

$$V=H^1_0(\Omega), \qquad A_0=-\Delta,$$

and $f(y)(z)=h(y(z))$ for every $z\in\Omega$, where h is a smooth (C^2_{Lip} at least) function $\mathbf{R}\to\mathbf{R}$ such that for some constants a, $b>0$ we have

$$\left|\frac{d^ih(w)}{dw^i}\right|\leqq a|w|^{3-i}+b \quad \text{for } i=0,1,2$$

($h(w)=w^3$, for instance). Finally let J be as in (7.2). Thus all our fundamental assumptions are verified. In particular, $f\in C^2_{\mathrm{Lip}}(H,H)$, because $H^1_0(\Omega)\subset L^6(\Omega)$.

REFERENCES

[1] V. BARBU, *Hamilton-Jacobi equations and nonlinear control problems*, J. Math. Anal. Appl., 120 (1986), pp. 494-509.

[2] V. BARBU AND G. DA PRATO, *Hamilton-Jacobi Equations in Hilbert Spaces*, Pitman, London, 1983.

[3] ———, *Hamilton-Jacobi equations in Hilbert spaces; variational and semigroup approach*, Ann. Mat. Pura Appl., 142 (1985), pp. 303-349.

[4] V. BARBU, G. DA PRATO, AND C. POPA, *Existence and uniqueness of the dynamic programming equation in Hilbert space*, Nonlinear Anal. Theory Meth. Appl., 7 (1983), pp. 283-299.

[5a] M. G. CRANDALL AND P. L. LIONS, *Hamilton-Jacobi equations in infinite dimensions. Part* I: *uniqueness of viscosity solutions*, J. Funct. Anal., 62 (1985), pp. 379-396.

[5b] ———, *Hamilton-Jacobi equations in infinite dimensions. Part* II: *existence of viscosity solutions*, J. Funct. Anal., 65 (1986), pp. 368-405.

[5c] ———, *Hamilton-Jacobi equations in infinite dimensions. Part* III, J. Funct. Anal., 68 (1986), pp. 214-247.

[5d] ———, *Hamilton-Jacobi equations in infinite dimensions. Part* IV: *Hamiltonians with unbounded linear terms*, J. Funct. Anal., 90 (1990), pp. 237-283.

[6] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Mathematics, Vol. 840, Springer-Verlag, Berlin, New York, 1981.

[7] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, Wiesbaden, 1972.

[8] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, Heidelberg, Berlin, 1983.

[9] Y. C. YOU, *A nonquadratic Bolza problem and a quasi-Riccati equation for distributed parameter systems*, SIAM J. Control Optim., 25 (1987), pp. 905-920.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 769-785, July 1991

MODEL MATCHING AND FACTORIZATION FOR NONLINEAR SYSTEMS: A STRUCTURAL APPROACH*

C. H. MOOG†, A. M. PERDON‡, AND G. CONTE§

Abstract. The model matching and the left factorization problems for nonlinear systems are investigated using an approach based on the structural algorithm. Sufficient conditions for the solvability of the first and necessary conditions, which in some cases and from a local point of view are also sufficient, for that of the second problem are found in terms of equalities between ranks or structures at infinity.

Key words. nonlinear systems, model matching, factorization, structure algorithm

AMS(MOS) subject classifications. 93B50, 93C05, 93C10

1. Introduction.

The problem of matching the input/output behavior of a model by precompensating a given system has been studied in the linear case by various authors and many results on the existence and the construction of general or special solutions are, by now, available (see [8], [9], [12], [20], [22], [23], and [32]). In the framework of nonlinear systems, the model matching problem was first considered in a general form in [7]. In that paper, the desired precompensator was required to be proper and the problem was tackled by reducing it to an equivalent disturbance decoupling problem. Two conditions that are respectively sufficient and necessary for the existence of solutions were obtained by using differential geometric techniques (see [15]). Then, extending the results of [20] such conditions were expressed in terms of equality of the structure at infinity, in the sense of [26], of the system and of a suitable composition of the system and the model. The case in which the model is a linear system was investigated in [16] and more recently in [4]. Necessary and sufficient conditions for a restricted class of nonlinear systems have recently been given in [14]. Other contributions are found in [27] and, from a differential algebraic point of view, in [28]. Remaining in a linear context, the model matching problem can be assimilated to that of factoring the transfer function of the model through that of the system, viewed as a left factor. The problem of factoring through a given right factor, which is called left factorization problem, can therefore be viewed as a sort of dual of the model matching one and, although less relevant in control theory, has received some attention. Results, which parallel those concerning the matching, are found in [1], [2], [12], [19], and [32].

In this paper we consider the model matching and the left factorization problems for affine nonlinear systems in a general formulation, which does not demand the compensator or the left factor to be proper, as well as in a stronger one, in which the properness is required. Our approach is based on the structure algorithm that is described in [18] and [31]. The algorithm is used as a tool both for defining the

* Received by the editors August 3, 1988; accepted for publication (in revised form) June 27, 1990.

† Laboratoire d'Automatique de Nantes, Unitée Associée au Centre National de la Recherche Scientifique, Ecole Nationale Supérieure de Mécanique, 1 rue de la Noë, 44072 Nantes Cedex 03, France.

‡ Dipartimento Metodi e Modelli Matematici per le Scienze Applicate, Università di Padova, Via Belzoni 7, 35100 Padova, Italy. This work was performed while the author was visiting the Laboratoire d'Automatique de Nantes with the support of Centre National de la Recherche Scientifique and partial support of Ministero della Pubblica Istruzione (fondi 40%).

§ Dipartimento di Matematica, Università di Genova, Via L. B. Alberti 4, 16132 Genova, Italy. This work was performed while the author was visiting the Laboratoire d'Automatique de Nantes with the support of Ministère de l'Education Nationale and partial support of Ministero della Pubblica Istruzione (fondi 40%).

structural invariants, rank, and structure at infinity, which allow us to state conditions for the existence of solutions, and for constructing such solutions. The results we obtain can be summarized as follows. The model matching problem admits a solution, respectively, a proper one, if the rank of the system and that of a suitable composition of the system and the model are equal, respectively, if their structure at infinity are equal. Note that the structure at infinity we refer to is that derived in [21] from the structure algorithm and does not coincide with that defined, in a geometric way, in [26] and considered in [7]. The sufficient condition we obtain is weaker than that of [7], however it is not necessary, as shown by an example due to Huijberts [13]. A nonstructural weaker condition is briefly discussed. The proof of the existence of a solution, when the sufficient condition is verified, is constructive. The results holding for the left factorization problem are the following. The problem admits a solution, respectively, a proper one, only if the rank of the factor system and that of the parallel composition of the same and the system to be factored are equal, respectively, if their structure at infinity are equal. Conversely, the rank equality is not in general sufficient and additional technical conditions are required to show, only locally and making use of the implicit function theorem, the existence of solutions.

The paper is organized as follows. Section 2 contains some preliminaries concerning the structure algorithm, the structure invariants defined by it, and their characterization in terms of vector spaces of differentials [5], [6]. Section 3 contains the statement and a discussion of the model matching problem, the related results that we already mentioned, and some examples. Section 4 deals with the left factorization problem, presenting and discussing its statement, the related results, and one example.

2. Notation and preliminaries. Consider a nonlinear control system of the form

$$\Sigma=\begin{cases}\dot{x}=f(x)+g(x)u, & \text{(2.1a)}\\ y=h(x), & \text{(2.1b)}\end{cases}$$

where $x(t)\in\mathbb{R}^n$, $u(t)\in\mathbb{R}^m$, $y(t)\in\mathbb{R}^p$, and $f(\cdot)$, the columns of $g(\cdot)$ and $h(\cdot)$, are *meromorphic* functions of x; that is, they are elements of the fraction field $\mathcal{F}$ of the ring of functions of the variable x, which are analytic on a domain $\mathcal{D}\subset\mathbb{R}^n$.

Following [5] and [6], we associate to $\sum$ a chain of vector spaces over the field $\mathcal{K}$ of meromorphic functions of $x, u, \cdots, u^{(n-1)}$ defined as follows. Recall, first, that denoting by $\nu=(\nu_1,\cdots,\nu_j)$ the components of $(x, u, \cdots, u^{(n-1)})$, the action of the partial derivative operator $\partial/\partial\nu_i$ on a meromorphic function $\eta(\nu)=\pi(\nu)/\theta(\nu)$, where $\pi(\cdot)$ and $\theta(\cdot)$ are analytic, is defined, by the usual rule of calculus, as

$$\frac{\partial}{\partial\nu_i}\frac{\pi(\nu)}{\theta(\nu)}:=\left(\theta(\nu)\frac{\partial}{\partial\nu_i}\pi(\nu)-\pi(\nu)\frac{\partial}{\partial\nu_i}\theta(\nu)\right)\Big/\theta^2(\nu). \tag{2.2}$$

Then, the differential of η is given by

$$d\eta(\nu):=\frac{\partial\eta}{\partial\nu}d\nu. \tag{2.3}$$

Returning to the system $\sum$, we view the time derivatives of the output y

$$\dot{y}=\dot{y}(x,u)=\frac{\partial y}{\partial x}[f(x)+g(x)u], \tag{2.4a}$$

$$\vdots$$

$$y^{(k+1)}=y^{(k+1)}(x,u,\cdots,u^{(k)})$$

$$=\frac{\partial y^{(k)}}{\partial x}[f(x)+g(x)u]+\sum_{i=0}^{k-1}\frac{\partial y^{(k)}}{\partial u^{(i)}}u^{(i+1)} \tag{2.4b}$$

as functions of x, $u, \cdots, u^{(k)}$, whose components are elements of $\mathcal{K}$. Now we can introduce the vector space $\mathscr{E}$ spanned over $\mathcal{K}$ by $\{dx, du, \cdots, du^{(n-1)}\}$ and we consider the chain of subspaces $\mathscr{E}_0 \subset \mathscr{E}_1 \subset \cdots \subset \mathscr{E}_n$ of $\mathscr{E}$ defined by

$$\begin{aligned} &\mathscr{E}_0 = \operatorname{span}\{dx\}, \\ &\qquad\vdots \\ &\mathscr{E}_n = \operatorname{span}\{dx, d\dot{y}, \cdots, dy^{(n)}\}. \end{aligned} \tag{2.5}$$

The notation dx stands for $dx_1, \cdots, dx_n$; du stands for $du_1, \cdots, du_m$, and $dy^{(k)}$ stands for $dy_1^{(k)}, \cdots, dy_p^{(k)}$ for any $k \geqq 0$.

Associated with the chain $\mathscr{E}_0 \subset \mathscr{E}_1 \subset \cdots \subset \mathscr{E}_n$ we have the list of integers $\sigma_1 \leqq \cdots \leqq \sigma_n$ given by

$$\sigma_k = \dim_{\mathcal{K}} \frac{\mathscr{E}_k}{\mathscr{E}_{k-1}}. \tag{2.6}$$

It has been shown in [5] and [6] that σ_1 equals the integer ρ_1, and $(\sigma_i - \sigma_{i-1})$ equals the integer ρ_i for $i \geqq 1$, where ρ_i is obtained from the ith step of the structure algorithm, or Singh's inversion algorithm [18], [31]. In particular, σ_n equals the rank ρ of $\sum$ as well as, in a suitable context, the differential output rank [10], [11]. It is useful, for completeness, to recall here Singh's inversion algorithm in a special form that will be employed in the sequel. Given the system $\sum$ we may consider its input u as divided into two subsets $u = (v, w)$, where v is viewed as a set of controls and w as a set of parameters. In this case, we apply to $\sum$ the following algorithm, denoted by Singh$_v$

Step 1. Calculate

$$\dot{y} = \frac{\partial h}{\partial x}[f(x) + g_v(x)v + g_w(x)w] =: f_1(x, w) + g_1(x)v \tag{2.7}$$

and set $G_1(x) := g_1(x)$ and $s_1 := \operatorname{rank} G_1(x)$. Permute, if necessary, the rows of the output so that the first s_1 rows of $G_1(x)$ are linearly independent and decompose $\dot{y}$ as

$$\dot{y} = \begin{pmatrix} \tilde{y}_1 \\ \hat{y}_1 \end{pmatrix}, \tag{2.8}$$

where $\dim \tilde{y}_1 = s_1 =: \rho_{1v}$. Then, eliminating v in the last rows, write

$$\begin{pmatrix} \tilde{y}_1 \\ \hat{y}_1 \end{pmatrix} = \begin{pmatrix} \tilde{f}_1(x, w) + \tilde{g}_1(x)v \\ \hat{y}_1(x, w, \tilde{y}_1) \end{pmatrix}$$

and set $\tilde{G}_1(x) =: \tilde{g}_1(x)$.

Step $k+1$. Suppose that from Steps 1 through k we have

$$\begin{aligned} \tilde{y}_1 &= \tilde{f}_1(x, w) + \tilde{g}_1(x)v, \\ &\ \ \vdots \\ \tilde{y}_k &= \tilde{f}_k(x, w, \cdots, w^{(k-1)}, \tilde{y}_1, \cdots, \tilde{y}_1^{(k-1)}, \cdots, \tilde{y}_{k-1}, \dot{\tilde{y}}_{k-1}) \\ &\qquad + \tilde{g}_k(x, w, \cdots, w^{(k-2)}, \tilde{y}_1, \cdots, \tilde{y}_1^{(k-2)}, \cdots, \tilde{y}_{k-1})v, \\ \hat{y}_k &= \hat{y}_k(x, w, \cdots, w^{(k-1)}, \tilde{y}_1, \cdots, \tilde{y}_1^{(k-1)}, \cdots, \tilde{y}_k), \end{aligned}$$

where

$$\tilde{G}_k = \begin{pmatrix} \tilde{g}_1 \\ \vdots \\ \tilde{g}_k \end{pmatrix}$$

has full rank s_k. Then

$$\begin{aligned}\dot{\hat{y}}_k = & f_{k+1}(x, w, \cdots, w^{(k)}, \tilde{y}_1, \cdots, \tilde{y}_1^{(k)}, \cdots, \tilde{y}_k, \dot{\tilde{y}}_k) \\ & + g_{k+1}(x, w, \cdots, w^{(k-1)}, \tilde{y}_1, \cdots, \tilde{y}_1^{(k-1)}, \cdots, \tilde{y}_k)v.\end{aligned}$$

Define $G_{k+1} := \left(\begin{smallmatrix}\tilde{G}_k \\ g_{k+1}\end{smallmatrix}\right)$ and $s_{k+1} := \text{rank } G_{k+1}(x)$. Decompose $\dot{\hat{y}}_k$ as

$$\dot{\hat{y}}_k = \begin{pmatrix}\tilde{y}_{k+1} \\ \hat{y}_{k+1}\end{pmatrix},$$

where $\dim \tilde{y}_{k+1} = s_{k+1} - s_k =: \rho_{k+1v}$. Then, eliminating v in the last rows, write

$$\begin{aligned}\tilde{y}_{k+1} = & \tilde{f}_{k+1}(x, w, \cdots, w^{(k)}, \tilde{y}_1, \cdots, \tilde{y}_1^{(k)}, \cdots, \tilde{y}_k, \dot{\tilde{y}}_k) \\ & + \tilde{g}_{k+1}(x, w, \cdots, w^{(k-1)}, \tilde{y}_1, \cdots, \tilde{y}_1^{(k-1)}, \cdots, \tilde{y}_k)v, \\ \hat{y}_{k+1} = & \hat{y}_{k+1}(x, w, \cdots, w^{(k)}, \tilde{y}_1, \cdots, \tilde{y}_1^{(k)}, \cdots, \tilde{y}_{k+1})\end{aligned}$$

and set

$$\tilde{G}_{k+1} := \begin{pmatrix}\tilde{g}_1 \\ \vdots \\ \tilde{g}_k\end{pmatrix}.$$

The above algorithm performs the inversion of $\sum$, viewed as a system depending on the parameter w, with respect to the input v when $\rho_v := s_n$ equals the dimension of v. When w is empty Singh$_v$ reduces to the usual Singh's inversion algorithm.

The indices σ_i, s_i, and ρ_i contain essentially the same information and each of them could be used in the following. We choose to state the next results in terms of the ρ_i, which have a direct interpretation as numbers of zeros at infinity of order i (see [21]), although the other indices are often used in proofs and calculations.

LEMMA 2.1. *Let the systems*

$$T = \begin{cases}\dot{x} = f(x) + g(x)u, \\ y_T = h(x) + h'(x)u\end{cases} \tag{2.9}$$

and

$$G = \begin{cases}\dot{z} = f_G(z) + g_G(z)v, \\ y_G = h_G(z)\end{cases} \tag{2.10}$$

with outputs of the same dimension, be given, and let (GT) *denote the composite system*

$$(GT) = \begin{cases}\dot{x} = f(x) + g(x)u, \\ \dot{z} = f_G(z) + g_G(z)v, \\ y_{GT} = h(x) - h_G(z) + h'(x)u.\end{cases} \tag{2.11}$$

Then we have $\rho_{iv}(GT) = \rho_i(G)$ *for all i and, in particular,* $\rho_v(GT) = \rho(G)$.

Proof. Let $\mathscr{K}'$ denote the field of meromorphic functions in the variables x, z, $v, \cdots, v^{(N-1)}$ and the parameters $u, \cdots, u^{(N)}$, where $N = \dim x + \dim z$. We denote by $\mathscr{E}_i^{GT}$ the vector space spanned over $\mathscr{K}'$ by $\{dx, dz, d\dot{y}_{GT}, \cdots, dy_{GT}^{(i)}\}$. Note that to consider $u, \cdots, u^{(N)}$ as parameters instead of variables means that the differential $d(\cdot)$ is given by $d(\cdot) = (\partial(\cdot)/\partial x)\, dx + (\partial(\cdot)/\partial z)\, dz + \sum_{i=0}^{N-1} (\partial(\cdot)/\partial v^{(i)})\, dv^{(i)}$. Following the proof given in Theorem 2.3 of [7] we can show that $\rho_{iv}(GT) = \dim_{\mathscr{K}'} \mathscr{E}_i^{GT}/\mathscr{E}_{i-1}^{GT}$. From this, since $dy_{GT}^{(j)} = dy_T^{(j)} - dy_G^{(j)} = \varphi_j(x, u, \cdots, u^{(j)})\, dx - dy_G^{(j)}$ with $\varphi_j \in \mathscr{K}'$, it follows that $\rho_{iv}(GT) = \dim \text{span}_{\mathscr{K}'} \{dx, dz, d\dot{y}_G, \cdots, dy_G^{(i)}\} - \dim \text{span}_{\mathscr{K}'} \{dx, dz, d\dot{y}_G,$

$\cdots, dy_G^{(i-1)}\}$, and hence $\rho_{iv}(GT) = \dim \operatorname{span}_{\mathcal{K}'}\{dz, d\dot{y}_G, \cdots, dy_G^{(i)}\} - \dim \operatorname{span}_{\mathcal{K}'}\{dz, d\dot{y}_G, \cdots, dy_G^{(i-1)}\}$. Now, let $\{\omega_1, \cdots, \omega_{r_i}\} \subset \{dz, d\dot{y}_G, \cdots, dy_G^{(i)}\}$ be a basis over $\mathcal{K}'$ of $\mathcal{E}_i^{GT}$, and let $\bar{\omega}$ be an element of $\{dz, d\dot{y}_G, \cdots, dy_G^{(i)}\} \backslash \{\omega_1, \cdots, \omega_{r_i}\}$. We write $\bar{\omega} = \sum \gamma_j(x, z, v, \cdots, v^{(N-1)}, u, \cdots, u^{(N)})\omega_j$ with $\gamma_j \in \mathcal{K}'$ and, computing the derivatives with respect to $x, u, \cdots, u^{(N)}$, we get

$$\frac{\partial \bar{\omega}}{\partial x} = \sum \frac{\partial \gamma_j}{\partial x} \omega_j = 0,$$

$$\frac{\partial \bar{\omega}}{\partial u} = \sum \frac{\partial \gamma_j}{\partial u} \omega_j = 0,$$

$$\vdots$$

$$\frac{\partial \bar{\omega}}{\partial u^{(N)}} = \sum \frac{\partial \gamma_j}{\partial u^{(N)}} \omega_j = 0.$$

Therefore $\partial\gamma_j/\partial x = 0$ and $\partial\gamma_j/\partial u = \cdots = \partial\gamma_j/\partial u^{(N)} = 0$ for all j, or, equivalently, $\gamma_j = \gamma_j(z, v, \cdots, v^{(N-1)})$. This says that $\{\omega_1, \cdots, \omega_{r_i}\}$ is a set of generators over the field $\mathcal{K}$ of meromorphic functions in the variables $z, v, \cdots, v^{(N-1)}$ of $\operatorname{span}_{\mathcal{K}}\{dz, d\dot{y}_G, \cdots, dy_G^{(i)}\} = \mathcal{E}_i^G$. Moreover, since $\mathcal{K} \subset \mathcal{K}'$, $\dim_{\mathcal{K}'} \mathcal{E}_i^{GT} = \dim_{\mathcal{K}} \mathcal{E}_i^G$ for all i and the result follows.

3. Model matching problem.

3. Model matching problem. In the nonlinear framework, the model matching problem was considered in [7] and, in the case of a linear model, in [4] and [16]. Some further contributions are made in [27]. The formulation of the model matching problem we give in the following differs slightly from that of [7].

3.1. Problem formulation. Let us now state the model matching problem (M.M.P.) in the formulation that will be used in the sequel.

MODEL MATCHING PROBLEM (M.M.P.). Given a model

$$T = \begin{cases} \dot{x} = f(x) + g(x)u, \\ y_T = h(x) \end{cases} \tag{3.1}$$

and a system G as in (2.10), find a proper compensator

$$H = \begin{cases} \dot{\xi} = f_H(\xi, z, u), \\ v = h_H(\xi, z, u) \end{cases}$$

with state space $\mathbb{R}^q$ and a map $\varphi: \mathbb{R}^n \to \mathbb{R}^q$ such that, denoting by y_{GH} the output of the composite system GH, we have that $y_T(u, x) - y_{GH}(u, \varphi(x), z)$; that is, the difference between the output of the model, viewed as a function of u and of the initial state x, and the output of the composite system, viewed as a function of u and of the initial states z and $\xi = \varphi(x)$, does not depend on u.

In order to gain a better insight into the model matching problem we are considering, we now state it in a generalized form (G.M.M.P.), which includes in particular the left inversion problem. Specializing such formulation by requiring a proper compensator we get the most interesting case from the point of view of control theory.

GENERALIZED MODEL MATCHING PROBLEM (G.M.M.P.). Given a model T as in (2.9) and a system G as in (2.10) find an integer $\nu \geqq 0$, a possibly nonproper compensator

$$H=\begin{cases}\dot{\xi}=f_H(\xi,z,u,\cdots,u^{(\nu)}),\\ v=h_H(\xi,z,u,\cdots,u^{(\nu)})\end{cases}\tag{3.2}$$

with state space $\mathbb{R}^q$ and a map $\varphi:\mathbb{R}^n\to\mathbb{R}^q$ such that, denoting by y_{GH} the output of the composite system GH, we have that $y_T(u,x)-y_{GH}(u,\varphi(x),z)$; that is, the difference between the output of the model, viewed as a function of u and of the initial state x, and the output of the composite system, viewed as a function of u and of the initial states z and $\xi=\varphi(x)$, does not depend on $u^{(\nu)}$.

The M.M.P. is the special case of the G.M.M.P. for $\nu=0$.

Remark 3.2. (i) In the M.M.P. the requirement that $y_T(u,x)-y_{GH}(u,\varphi(x),z)$ does not depend on u amounts, in the linear case, to the equality of the transfer functions of the model and of the composite systems. From this point of view, therefore, our formulation represents the natural extension of the one currently understood for the linear model matching problem (compare with the references quoted in the Introduction and with [15] and [16]).

We recall that a stronger formulation of the M.M.P., requiring the equality of y_T and y_{GH}, has been considered, only for a linear model, in [4]. Note that the problem we stated qualifies as an exact M.M.P., as opposed to an approximate or an asymptotic M.M.P., which could also be considered (see, e.g., [14]).

Let us consider the left inversion problem in the linear framework. The solution provided by the Silverman algorithm [30] has the form (3.2), where ν is the inherent integration order of the system [29], [25]. In the simple example given by $T=\{y_T=y$ and by

$$G=\begin{cases}\dot{z}=v,\\ y=z,\end{cases}$$

we obtain

$$H=G^{-1}=\begin{cases}\dot{\xi}=\dot{y},\\ v=\dot{y}.\end{cases}$$

The difference between the outputs of the identity model T and of GH is $y_T-y_{GH}=y-z$; the latter depends on the input y and is independent on the first derivative $\dot{y}$.

(ii) Roughly speaking, our formulation of the G.M.M.P. amounts to requiring that $y_T(u,x)-y_{GH}(u,\varphi(x),z)$ depend only on a finite number of derivatives of the input. In our framework, the independence of $y_T(u,x)-y_{GH}(u,\varphi(x),z)$ on $u^{(\nu)}$ is technically expressed by the following condition:

$$d(y_T(u,x)-y_{GH}(u,\varphi(x),z))^{(k)}\in\operatorname{span}_{\mathcal{K}}\{dx,dz,d\xi,du,\cdots,du^{(\nu-1)}\}\quad\text{for } k\geqq 0,\tag{3.3}$$

which says, in other terms, that $u^{(\nu)}$ does not appear in any time derivative of $y_T(u,x)-y_{GH}(u,\varphi(x),z)$. In (3.3), $\mathcal{K}$ denotes the field of meromorphic functions in the variables $x, z, \xi, u, \cdots, u^{(\nu+N)}$ with $N=\dim x+\dim z+\dim\xi$.

(iii) In dealing with nonlinear systems of the form (2.1), we must take into account the presence of possible singularities. To cope with them, we will say that the M.M.P. is solvable if there exists a compensator H of the form (3.2) and a map φ, which achieve (3.3) for all initial states x and z in an open and dense subset of the state spaces and for all (sufficiently many times differentiable) input functions $u(t)$, assuming values in an open dense subset of $\mathbb{R}^n$, while the system evolve on a time interval $[0, t)$, whose length may depend on the chosen initial conditions and on the input. In such a case the pair (H,φ) is called a solution of the M.M.P. The next examples will illustrate our point of view.

Example 3.3. (i) Let

$$T = \begin{cases} \dot{x} = u, \\ y_T = x \end{cases} \quad \text{and} \quad G = \begin{cases} \dot{z} = v, \\ y_G = z^2 \end{cases}$$

be the data of a M.M.P. The pair consisting of the compensator $H = \{v = u/2z$ and of the empty function is a solution in the sense of Remark 3.2. In fact, for $z_0 \neq 0$, we obtain $y_{GH}(u, z_0) = \int_0^t u(\tau)\, d\tau + z_0^2$ for all input functions $u(\tau)$ and for all $t > 0$ such that $\int_0^t u(\tau)\, d\tau + z_0^2 > 0$. Then, we have $y_T(u, x) - y_{GH}(u, z) = x - z^2$ and $d(y_T(u, x) - y_{GH}(u, z))^{(k)} \in \operatorname{span}_{\mathcal{K}} \{dx, dz\}$. In particular, if, for example, the input is bounded by $|u(\tau)| \leqq M$, and the initial conditions $x_0, z_0 \neq 0$, are chosen, $y_T - y_{GH}$ is independent from u over the time interval $[0, z_0^2/M)$. It may be useful to note that $v = u/2z$ is a solution of the M.M.P. in the same way, that is with the same limitations, in which it is a solution, in the sense of [17], of the disturbance decoupling problem with disturbance measurement described by $\dot{x} = u$, $\dot{z} = v$, $y = x - z^2$, where u is the disturbance and v is the control.

(ii) Note that taking, for instance,

$$T = \begin{cases} \dot{x} = xu, \\ y_T = x \end{cases} \quad \text{and} \quad G = \begin{cases} \dot{z} = zv, \\ y_G = z, \end{cases}$$

contrary to what happens in the linear case, the identity compensator $H = \{v = u$ does not give a solution of the M.M.P. In fact, $y_T - y_{GH} = (x_0 - z_0) \exp (\int_0^t u(\tau)\, d\tau)$ is independent from u only if the initial states of the model and of the system coincide. In this case a solution is given by the compensator

$$H = \begin{cases} \dot{\xi} = \xi u, \\ v = (\xi/z) u, \end{cases} \quad \text{where } \xi(t) \in \mathbb{R}^n,$$

and by $\varphi = \mathrm{id}$.

A structural condition under which a compensator exists and a procedure to compute it are given in the following theorem.

THEOREM 3.4. *The generalized model matching problem is solvable if*

$$\rho(GT) = \rho(G), \tag{3.4}$$

where (GT) *is the composite system* (2.11).

Proof. Applying Singh$_v$ to (GT), we obtain

$$\begin{aligned} \begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \\ \vdots \\ \tilde{Y}_N \\ \hat{Y}_N \end{pmatrix} &= \begin{pmatrix} \tilde{F}_1(x, z, u, \dot{u}) \\ \tilde{F}_2(x, z, u, \dot{u}, \ddot{u}, \tilde{Y}_1, \dot{\tilde{Y}}_1) \\ \vdots \\ \tilde{F}_N(x, z, u, \cdots, u^{(N)}, \tilde{Y}_1, \cdots, \tilde{Y}_1^{(N-1)}, \cdots, \tilde{Y}_{N-1}, \dot{\tilde{Y}}_{N-1}) \\ \hat{F}_N(x, z, u, \cdots, u^{(N)}, \tilde{Y}_1, \cdots, \tilde{Y}_1^{(N-1)}, \cdots, \tilde{Y}_{N-1}, \dot{\tilde{Y}}_{N-1}, \tilde{Y}_N) \end{pmatrix} \\ &\quad + \begin{pmatrix} \tilde{G}_1(z) \\ \tilde{G}_2(x, z, u, \tilde{Y}_1) \\ \vdots \\ \tilde{G}_N(x, z, u, \cdots, u^{(N-2)}, \tilde{Y}_1, \cdots, \tilde{Y}_1^{(N-2)}, \cdots, \tilde{Y}_{N-1}) \\ 0 \end{pmatrix} v \\ &= \begin{pmatrix} \tilde{F} \\ \hat{F}_N \end{pmatrix} + \begin{pmatrix} \tilde{G} \\ 0 \end{pmatrix} v \end{aligned} \tag{3.5}$$

with rank $\tilde{G} = \#$ rows $\tilde{G} = \rho_v(GT)$ and where $\tilde{Y}_i$ represents a suitable subset of rows of $y^{(i)}_{GT}$, which will be useful to denote also as $y^{(i)}_{T.i} - y^{(i)}_{G.i}$. We can choose constant values

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_{N-1} \\ 0 \end{pmatrix} \quad \text{for } \tilde{Y} = \begin{pmatrix} \tilde{Y}_1^{(N-2)} \\ \vdots \\ \tilde{Y}_{N-1} \\ \tilde{Y}_N \end{pmatrix}$$

such that the generic rank of $\tilde{G}$ evaluated at Y is equal to the number of rows of $\tilde{G}$. Then, solving for v the system

$$\begin{pmatrix} \tilde{Y}_1 \\ \vdots \\ \tilde{Y}_N \end{pmatrix}_{|Y} = \tilde{F}_{|Y} + \tilde{G}_{|Y} v,$$

obtained by replacing $\tilde{Y}$ with Y in (3.5), we get

$$v = \phi(x, z, u, \cdots, u^{(N)}, \tilde{Y}_1, \cdots, \tilde{Y}_1^{(N-3)}, Y_1, \cdots, \tilde{Y}_{N-2}, Y_{N-2}, Y_{N-1}).$$

Now, denoting by w_0 a vector of same dimension as x and by w_i a vector of dimension $(N-i)\cdot \dim \tilde{Y}_i$, we set $\nu = N$ and we construct the compensator

$$(3.6) \qquad H = \begin{cases} \dot{w}_0 = f(w_0) + g(w_0)u \\ \dot{w}_i = \begin{pmatrix} 0 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 \\ 0 & \cdot & \cdot & 0 \end{pmatrix} w_i + \begin{pmatrix} 0 \\ \vdots \\ 0 \\ Y_i \end{pmatrix} & \text{for } 1 \leqq i \leqq N-2, \\ v = \phi(w_0, z, u, \cdots, u^{(N)}, w_1, \cdots, w_{N-2}, Y_{N-1}) \end{cases}$$

and, letting $\varphi(x) = (x, 0, \cdots, 0)$, we claim that (H, φ) is a solution of the G.M.M.P. In order to show this, let us first remark that, in (3.5), $\hat{Y}_N$ is actually independent from $u^{(k)}$, for all k. In fact, by Lemma 2.1 and the rank equality (3.4) it follows that $\rho(GT) = \rho_v(GT)$ and we know from Theorem 2.3 of [7], that the dx, dz, $d\tilde{Y}_1, \cdots, d\tilde{Y}_1^{(N-1)}, \cdots, d\tilde{Y}_N$ are independent over the field $\mathcal{K}$. So, if $\partial \hat{F}_N / \partial u^{(k)} \neq 0$, for some $k \geqq 0$, $d\hat{Y}_N$ does not belong to $\text{span}_{\mathcal{K}}\{dx, dz, d\tilde{Y}_1, \cdots, d\tilde{Y}_1^{(N-1)}, \cdots, d\tilde{Y}_N\}$, then $\rho(GT) > \rho_v(GT)$, contradicting the assumption.

Now let us consider the composite system (GH):

$$(GH) = \begin{cases} \dot{w} = F(w) + G(w)u, \\ \dot{z} = f_G(z) + g_G(z)\phi(w, z, u, \cdots, u^{(N)}), \\ y_{GH} = h_G(z) \end{cases}$$

initialized at $\varphi(x_0) = (x_0, 0, \cdots, 0)$, and the difference $y_T - y_{GH}$ between the output of the model and that of (GH). Recalling the notation $\tilde{Y}_i = y^{(i)}_{T.i} - y^{(i)}_{G.i}$, by substituting the output of H to v in (3.5) and taking derivatives, we get

$$y^{(N-1)}_{T.i} - y^{(N-1)}_{GH.i} = Y_i \quad \text{for } 1 \leqq i \leqq N-1,$$
$$y^{(N)}_{T.N} - y^{(N)}_{GH.N} = 0.$$

Therefore $d(y_T - y_{GH})^{(k)} \in \text{span}_{\mathcal{K}}\{dx, dz, dw, du, \cdots, du^{(N-1)}\}$ for all k. □

It is worthwhile to note that, although in § 3.1 the compensator H is described in a very general form, the construction illustrated in the proof of Theorem 3.4 always produces a system whose state equations have the same form of those of the model.

In particular, the derivatives of the input appear only in the output function $h_H(\xi, z, u, \cdots, u^{(\nu)})$. A structural condition under which there exists a proper compensator H, that is, one that does not depend on the derivatives of the input u, is given in the next theorem.

THEOREM 3.5. *The* M.M.P. *is solvable with a proper compensator H of the form*

$$H=\begin{cases}\dot{\xi}=f_H(\xi, z, u),\\ v=h_H(\xi, z, u)\end{cases}$$

if

$$\rho_i(GT)=\rho_i(G) \tag{3.7}$$

for all $i \geqq 1$.

Proof. Assume that (3.7) holds; then, for all i, we have by Lemma 2.1 $\rho_{iv}(GT)=\rho_i(GT)$. In particular, this implies that at the first step of the algorithm Singh$_v$ applied to (GT) we have

$$\begin{pmatrix}\tilde{Y}_1\\ \hat{Y}_1\end{pmatrix}=\begin{pmatrix}\tilde{F}_1(x, z, u)\\ \hat{F}_1(x, z, u, \tilde{Y}_1)\end{pmatrix}+\begin{pmatrix}\tilde{G}_1(z)\\ 0\end{pmatrix}v,$$

where actually $\partial\hat{F}_1/\partial u=0$, since otherwise $\rho_1(GT)$ would be strictly greater than $\rho_{1v}(GT)$. Repeatedly applying the same argument, we get at the last step N

$$\begin{pmatrix}\tilde{Y}\\ \hat{Y}_N\end{pmatrix}=\begin{pmatrix}\tilde{F}(x, z, u, \tilde{Y}, \cdots, \tilde{Y}^{(N-1)})\\ \hat{F}_N(x, z, u, \tilde{Y}, \cdots, \tilde{Y}^{(N)})\end{pmatrix}+\begin{pmatrix}\tilde{G}(x, z, \tilde{Y}, \cdots, \tilde{Y}^{(N-2)})\\ 0\end{pmatrix}v$$

and, hence, $v=\phi(x, z, u, \tilde{Y}, \cdots, \tilde{Y}^{(N-3)})$. Therefore the compensator obtained following the construction described in the proof of Theorem 3.4 is, in this case, proper. □

Example 3.6. (i) The M.M.P. concerning the model

$$T=\begin{cases}\dot{x}=\begin{pmatrix}x_2\\ 0\\ x_4\\ 0\end{pmatrix}+\begin{pmatrix}0&0\\ 1&0\\ 0&0\\ 0&1\end{pmatrix}u,\\ y_T=\begin{pmatrix}x_1\\ x_3\end{pmatrix}\end{cases}$$

and the system

$$G=\begin{cases}\dot{z}=\begin{pmatrix}0\\ z_4\\ 0\\ 0\end{pmatrix}+\begin{pmatrix}z_3&0\\ 0&1\\ 1&1\\ 0&1\end{pmatrix}v,\\ y_G=\begin{pmatrix}z_2-z_3\\ z_1\end{pmatrix}\end{cases}$$

was considered in [7]. It was shown that the geometric necessary condition given in the same paper is not verified, although the compensator

$$H=\begin{cases}\dot{\zeta}=\dfrac{\zeta^2}{\zeta+z_3}+\dfrac{1}{\zeta+z_3}(-\zeta\quad 1)u,\\ v=\begin{pmatrix}\zeta\\ \zeta^2/(\zeta+z_3)\end{pmatrix}+\dfrac{1}{\zeta+z_3}\begin{pmatrix}0&0\\ z_3&1\end{pmatrix}u\end{cases}$$

provides a solution of the problem (see [7, Ex. 5.4]). It can be easily checked that (3.7) is verified, that is, $\rho_i(GT)=\rho_i(G)$. Then, applying the procedure illustrated in the proof of Theorem 3.4, we get the proper compensator

$$H'=\begin{cases} \dot{\xi}=\begin{pmatrix}\xi_2\\0\\\xi_4\\0\end{pmatrix}+\begin{pmatrix}0&0\\1&0\\0&0\\0&1\end{pmatrix}u,\\ v=\begin{pmatrix} z_4-\xi_2\\ (z_4-\xi_2-u_1z_3-u_2)/(\xi_2-z_3-z_4)\end{pmatrix}.\end{cases}$$

Clearly, by removing the unnecessary equations $\dot{\xi}_1=\xi_2$, $\dot{\xi}_3=\xi_4$, and $\dot{\xi}_4=u_2$, we obtain another compensator, say H'', which solves the problem. Now, the change of variables $\zeta=z_4-\xi_2$ transforms H'' into H.

(ii) Consider the model

$$T=\begin{cases}\dot{x}=\begin{pmatrix}x_2\\0\end{pmatrix}+\begin{pmatrix}0\\1\end{pmatrix}u,\\ y_T=\begin{pmatrix}0\\x_1\end{pmatrix}\end{cases}$$

and the system

$$G=\begin{cases}\dot{z}=\begin{pmatrix}1&0\\0&1\\z_3&1\end{pmatrix}v,\\ y_G=\begin{pmatrix}z_1\\z_2-z_3\end{pmatrix},\end{cases}$$

for which $\rho(GT)=\rho(G)$. Note that, since

$$y_G=\begin{pmatrix}\displaystyle\int_0^t v_1(\tau)\,d\tau+z_1(0)\\ \displaystyle\int_0^t v_2(\tau)\,d\tau-\exp\left(\int_0^t v_1(\tau)\,d\tau\right)\int_0^t \exp\left(\int_0^\tau v_1(\sigma)\,d\sigma\right)v_2(\tau)\,d\tau\\ \displaystyle +z_2(0)+z_3(0)\exp\left(\int_0^t v_1(\tau)\,d\tau\right)\end{pmatrix}$$

and $y_{T.1}=0$, contrary to what happens in the linear case, it is not possible to find a compensator H such that $y_T-y_{GH}=0$ for $u\neq 0$, also when we are allowed to choose the initial condition $z(0)$. Applying Singh$_v$ to (GT), we get

$$\tilde{Y}=\begin{pmatrix}\dot{Y}_1\\\ddot{Y}_2\end{pmatrix}=\begin{pmatrix}0\\u-z_3\ddot{Y}_1\end{pmatrix}-\begin{pmatrix}1&0\\z_3\dot{Y}_1&\dot{Y}_1\end{pmatrix}v=\tilde{F}+\tilde{G}v,$$

and then, in fixing constant values Y for $\tilde{Y}$, we are obliged to choose $\dot{Y}_1\neq 0$. Taking, for instance, $Y=\binom{1}{0}$, we get $v=\binom{-1}{u+z_3}$, which represents by itself a compensator H that solves the problem.

Remark 3.7. The conditions of Theorems 3.4 and 3.5 are not necessary for the existence of solutions to the G.M.M.P. and the M.M.P. as pointed out by the following example due to Huijberts [13]. Let

$$T=\begin{cases} \dot{x}=\begin{pmatrix} x_2\\ x_3\\ 0\\ x_4 \end{pmatrix}+\begin{pmatrix} 0 & 0\\ 1 & 0\\ 0 & 1\\ 0 & 0 \end{pmatrix}\begin{pmatrix} u_1\\ u_2 \end{pmatrix},\\ y_T=\begin{pmatrix} x_2\\ x_4\\ x_1 \end{pmatrix} \end{cases}$$

and

$$G=\begin{cases} \dot{z}=\begin{pmatrix} z_2\\ 0\\ 0 \end{pmatrix}+\begin{pmatrix} 0 & z_2\\ 1 & 0\\ 0 & 1 \end{pmatrix}\begin{pmatrix} v_1\\ v_2 \end{pmatrix},\\ y_G=\begin{pmatrix} z_2\\ z_3\\ z_1 \end{pmatrix}. \end{cases}$$

By applying Singh's algorithm to G, we get $\rho_1(G)=\rho_2(G)=\rho_3(G)=2$. The same procedure applied to (GT) gives

$$\begin{aligned} \dot{y}_{GT1}&=x_3+u_1-v_1,\\ \dot{y}_{GT2}&=x_4-v_2,\\ \dot{y}_{GT3}&=x_2+z_2(\dot{y}_{GT2}-x_4-1), \end{aligned}$$

and then

$$\ddot{y}_{GT3}=\dot{y}_{GT1}+z_2(\ddot{y}_{GT2}-x_4)-(\dot{y}_{GT1}-x_3-u_1)(\dot{y}_{GT2}-x_4).$$

So $\rho_1(GT)=2$, $\rho_2(GT)=3$ and the sufficient conditions of Theorems 3.4 and 3.5 are not satisfied. However, the compensator

$$H=\begin{cases} \dot{\xi}=\begin{pmatrix} \xi_2\\ \xi_3\\ 0 \end{pmatrix}+\begin{pmatrix} 0 & 0\\ 1 & 0\\ 0 & 1 \end{pmatrix}\begin{pmatrix} u_1\\ u_2 \end{pmatrix},\\ v_1=\xi_3+u_1,\\ v_2=0, \end{cases}$$

and $\varphi=\mathrm{id}$ give a solution to the M.M.P.

From (3.5) we get the equality

$$\begin{aligned} \tilde{Y}=\tilde{F}(x,z,u,\cdots,u^{(N)},\tilde{Y}_1,\cdots,\tilde{Y}_1^{(N-1)},\cdots,\tilde{Y}_{N-1},\dot{\tilde{Y}}_{N-1})&\\ +\tilde{G}_N(x,z,u,\cdots,u^{(N-2)},\tilde{Y}_1,\cdots,\tilde{Y}_1^{(N-2)},\cdots,\tilde{Y}_{N-1})v& \end{aligned}\tag{3.8}$$

and, by derivation of $\hat{Y}_N$, the equalities

$$\begin{aligned} \hat{Y}_N&=\hat{F}_N(x,z,u,\cdots,u^{(N)},\tilde{Y}_1,\cdots,\tilde{Y}_1^{(N-1)},\cdots,\tilde{Y}_{N-1},\dot{\tilde{Y}}_{N-1},\tilde{Y}_N),\\ \hat{Y}_N^{(n+q-N)}&=\hat{F}_{n+q}(x,z,u,\cdots,u^{(n+q)},\tilde{Y}_1,\cdots,\tilde{Y}_1^{(n+q-1)},\\ &\qquad\cdots,\tilde{Y}_{N-1},\cdots,\tilde{Y}_{N-1}^{(n+q-N+1)},\tilde{Y}_N,\cdots,\tilde{Y}_N^{(n+q-N)}), \end{aligned}\tag{3.9}$$

from which it can be understood that a weaker condition for the existence of solutions to the M.M.P. is, in particular, that there exists a vector of functions

$$Y(x, z) = \begin{pmatrix} Y_1(x, z) \\ \vdots \\ Y_N(x, z) \end{pmatrix}$$

such that

(i) $\partial Y^{(k)}/\partial u = 0$ for all k;

(ii) Substituting $Y(x, z)$ and its derivatives for

$$\tilde{Y} = \begin{pmatrix} \tilde{Y}_1 \\ \vdots \\ \tilde{Y}_N \end{pmatrix}$$

and its derivatives in $\tilde{G}$, the generic rank is equal to the number of rows;

(iii) Substituting $Y(x, z)$ and its derivatives for

$$\tilde{Y} = \begin{pmatrix} \tilde{Y}_1 \\ \vdots \\ \tilde{Y}_N \end{pmatrix}$$

and its derivatives in $\hat{F}_N, \cdots, \hat{F}_{n+q}$, respectively in $\tilde{F}$, all the coefficients of the monomials in $u, \cdots, u^{(n+q)}$, and respectively, all the coefficients of the monomials in $\dot{u}, \cdots, u^{(N)}$, are zero. Such a condition is verified in the example above for $Y(x, z) = \binom{0}{x_4}$.

4. Left factorization problem. It is well known [2], [12] that in the linear case (3.4) and (3.7) are necessary and sufficient conditions for solving the G.M.M.P. or the M.M.P., and also when no feedback connection between the state of the system G and the precompensator H is allowed. In such formulation, the linear G.M.M.P. amounts to the problem of factoring the transfer function of the model T through a possible left factor, represented by the transfer function of G. It is then, natural, from an abstract point of view, to also consider the dual problem, which consists of factoring the transfer function of T through a possible given right factor (see [1], [2], [12], and [19]). In the more general case we are considering, this leads to the following formulation for what we call the left factorization problem.

4.1. Problem formulation.

Left Factorization Problem (L.F.P.). Given a model T as in (3.1) and a system

$$(4.1) \qquad H = \begin{cases} \dot{z} = f_H(z) + g_H(z)u, \\ v = h_H(z), \end{cases}$$

find a proper compensator

$$G = \begin{cases} \dot{\xi} = f_G(\xi, v), \\ y_G = h_G(\xi, v) \end{cases}$$

with state space $\mathbb{R}^q$ and a map $\varphi : \mathbb{R}^n \to \mathbb{R}^q$ such that, denoting by Y_{GH} the output of the cascade GH, we have $y_T(u, x_0) - y_{GH}(u, \varphi(x_0), z_0) = 0$ for any initial states x_0, z_0.

Generalized Left Factorization Problem (G.L.F.P.). Given a model T as in (2.9) and a system H as in (4.1), find an integer $\nu \geqq 0$, a possibly nonproper compensator

$$(4.2) \qquad G = \begin{cases} \dot{\xi} = f_G(\xi, v, \cdots, v^{(\nu)}), \\ y_G = h_G(\xi, v, \cdots, v^{(\nu)}), \end{cases}$$

with state space $\mathbb{R}^q$, and a map $\varphi:\mathbb{R}^n\to\mathbb{R}^q$, such that, denoting by Y_{GH} the output of the cascade GH, we have

$$y_T(u, x_0)-y_{GH}(u, \varphi(x_0), z_0)=0 \tag{4.3}$$

for any initial states x_0, z_0.

Remark 4.2. The same considerations as in Remark 3.2 apply to the present situation. Therefore a solution (G, φ) will be one that achieves (4.3) for all initial states x_0 and z_0 in an open and dense subset of the state spaces.

The first result we have in this framework is the following theorem.

THEOREM 4.3. *The* G.L.F.P. *is solvable only if*

$$\rho\begin{pmatrix}T\\ H\end{pmatrix}=\rho(H), \tag{4.4}$$

where $\left(\begin{smallmatrix}T\\ H\end{smallmatrix}\right)$ *is the system consisting of the state and output equations of T and H.*

Proof. We start by proving the theorem under an additional technical assumption on the system H. Assume that the maximal regular controllability distribution $\mathcal{R}^*_H$ of H contained in ker dh_H is locally well defined, i.e., that the regularity conditions of [15, § 6.4] are satisfied. Denoting by $\mathcal{G}$ the distribution spanned by $g_H(z)$, we assume that the following holds:

$$\dim(\mathcal{G}\cap\mathcal{R}^*_H)=m-\rho(H). \tag{4.5}$$

Now, let the regular feedback $u=\alpha(z)+\beta(z)w$ be a "friend" of $\mathcal{R}^*_H$, and let us denote by $\left(\begin{smallmatrix}\tilde T\\ \tilde H\end{smallmatrix}\right)$ the system obtained by compensating $\left(\begin{smallmatrix}T\\ H\end{smallmatrix}\right)$ with $u=\alpha(z)+\beta(z)w$. By (4.5), the action of the feedback $u=\alpha(z)+\beta(z)w$ transforms H into the system $\tilde H$ that, up to a change of coordinates, is of the form [15] $\dot z_1=f_1(z_1)+g_1(z_1)\bar w_1$, $\dot z_2=f_2(z_1, z_2)+g_2(z_1, z_2)w$, $v=h_{\tilde H}(z_1)$, where $w=(\bar w_1, \bar w_2)$, $\bar w_1=(w_1,\cdots, w_\rho)$ and $\rho=\rho(H)$. Hence we have

$$\frac{\partial v^{(k)}(w, z_0)}{\partial w_i}=0 \tag{4.6}$$

for all $i\geqq\rho+1$ and all k. Moreover, if (G, φ) is a solution of the G.L.F.P., we have that the output trajectory $\bar Y(w, x_0, \varphi(x_0), z_0)$ of the cascade composition between $\left(\begin{smallmatrix}\tilde T\\ \tilde H\end{smallmatrix}\right)$ and the system

$$\tilde G=\begin{cases}\dot\xi=f_G(\xi, v,\cdots, v^{(\nu)}),\\ y=y_{\tilde T}-h_G(\xi, v,\cdots, v^{(\nu)}),\end{cases}$$

initialized at $(x_0, \varphi(x_0), z_0)$, is identically zero. This, together with (4.5), implies

$$\begin{aligned}\frac{\partial y^{(k)}(w, x_0, \varphi(x_0), z_0)}{\partial w_i}&=\frac{\partial y^{(k)}_{\tilde T}(w, x_0, z_0)}{\partial w_i}-\frac{\partial y^{(k)}_{G\tilde H}(w, x_0, \varphi(x_0), z_0)}{\partial w_i}\\ &=\frac{\partial y^{(k)}_{\tilde T}(w, x_0, z_0)}{\partial w_i}-\frac{\partial y^{(k)}_{G}(y_{\tilde H}(w, z_0), \varphi(x_0))}{\partial w_i}\\ &=\frac{\partial y^{(k)}_{\tilde T}(w, x_0, z_0)}{\partial w_i}-\frac{\partial y^{(k)}_{G}(y_{\tilde H}(w, z_0), \varphi(x_0))}{\partial y_{\tilde H}}\cdot\frac{\partial y_{\tilde H}(w, z_0)}{\partial w_i}\\ &=\frac{\partial y^{(k)}_{\tilde T}(w, x_0, z_0)}{\partial w_i}\\ &=0\quad\text{for all } i\geqq\rho+1 \text{ and all } k.\end{aligned}$$

Therefore, $\rho\left(\begin{smallmatrix}\tilde T\\ \tilde H\end{smallmatrix}\right)$ is not greater than $\rho(H)$ and, as a consequence, $\rho\left(\begin{smallmatrix}T\\ H\end{smallmatrix}\right)=\rho\left(\begin{smallmatrix}\tilde T\\ \tilde H\end{smallmatrix}\right)=\rho(H)$.

The general case can always be reduced to the previous one. In fact, if (4.5) does not hold, we can pick $\rho(H)$ independent output components of H that can be decoupled with a regular dynamic state feedback [10], [3], [24]. Then, the extended system H_E verifies $\dim(\mathcal{G}_E \cap \mathcal{R}^*_{H_E}) = m - \rho(H_E)$. Since any solution of the G.L.F.P. concerning T and H also solves that concerning T_E and H_E, and since the regular dynamic state feedback does not affect the rank of the systems, the conclusion follows from the first part. □

In general, (4.4) is not sufficient for the solvability of the G.L.F.P. However, under (4.5) and an additional technical condition, which essentially assures the possibility of locally expressing z as a function of the output and its derivatives, it is possible to get a local result. More precisely, it is possible, for any z_0 in an open and dense subset of the state space, to find a neighborhood $\mathcal{D}_0$ and to show the existence of a compensator G and a map φ that achieve (4.3) for $z \in \mathcal{D}_0$. In this case, we will say, that the L.F.P. is locally solvable.

THEOREM 4.4. *The* G.L.F.P. *is locally solvable if the following conditions hold*:

(i) $\rho\binom{T}{H} = \rho(H)$;

(ii) $\dim(\mathcal{G} \cap \mathcal{R}^*_H) = m - \rho(H)$;

(iii) $\sum_{i \geqq 0} (\rho(H) - s_i) = n$, *where* $n = \dim z$, $s_0 = 0$, *and the* s_i *are obtained by applying Singh's inversion algorithm to H.*

Proof. We consider a friend of $\mathcal{R}^*_H$, $u = \alpha(z) + \beta(z)w$, as in the proof of Theorem 4.3 and we use the notation introduced there. By the rank equality $\rho\binom{\tilde{T}}{\tilde{H}} = \rho\binom{T}{H} = \rho(H)$, since (4.6) holds, the input components w_i, with $i \geqq \rho + 1$, do not affect the output of $\binom{\tilde{T}}{\tilde{H}}$. By applying Singh's inversion algorithm to $\tilde{H}$ we get

$$\bar{w}_1 = \psi(z, v, \dot{v}, \cdots, v^{(\nu)}), \tag{4.7}$$

where ψ is a meromorphic function of its arguments and, in particular, it is defined for all z in an open dense subset of the state space. Moreover, using arguments such as those in [17, § 4], we can show that, by (iii), the Jacobian matrix

$$\frac{\partial}{\partial z} \begin{pmatrix} v - h_{\tilde{H}}(z) \\ \hat{v}_1 - \hat{v}_1(z, \tilde{v}_1) \\ \vdots \\ \hat{v}_N - \hat{v}_N(z, \tilde{v}_1, \cdots, \tilde{v}_1^{(N-1)}, \cdots, \tilde{v}_N) \end{pmatrix},$$

whose elements are obtained by applying Singh's algorithm to $\tilde{H}$, has rank n. Then, for any z_0 in an open and dense subset of the state space, there exists a neighborhood $\mathcal{D}_0$ of z_0 such that $z = \chi(v, \dot{v}, \cdots, v^{(\nu)})$ for $z \in \mathcal{D}_0$. By substituting in (4.7), we then get, $\bar{w}_1 = \bar{\psi}(v, \dot{v}, \cdots, v^{(\nu)})$. Now, writing the state equation of $\binom{\tilde{T}}{\tilde{H}}$ as $\dot{x} = f_1(x, z) + g_1(x, z)\bar{w}_1 + g_2(x, z)\bar{w}_2$, $\dot{z} = f_2(z) + g_3(z)w$, we can consider the system

$$G = \begin{cases} \dot{\xi} = f_1(\xi, \chi(v, \dot{v}, \cdots, v^{(\nu)})) + g_1(\xi, \chi(v, \dot{v}, \cdots, v^{(\nu)}))\bar{\psi}(v, \dot{v}, \cdots, v^{(\nu)}), \\ y_G = h(\xi), \end{cases}$$

$$= \begin{cases} \dot{\xi} = f_G(\xi, v, \dot{v}, \cdots, v^{(\nu)}), \\ y_G = h(\xi), \end{cases}$$

and we claim that (G, φ), where φ is the identity map, is a solution of the G.L.F.P. relatively to $\mathcal{D}_0$. In fact, by inspection, we see that the output trajectory

$Y(w, x_0, \varphi(x_0), z_0)$ of the system

$$\begin{aligned}
\dot{x} &= f_1(x, z) + g_1(x, z)\bar{w}_1 + g_2(x, z)\bar{w}_2,\\
\dot{z} &= f_2(z) + g_3(z)w,\\
\dot{\xi} &= f_1(\xi, z) + g_1(\xi, z)\bar{w}_1,\\
y &= h(x) - h(\xi)
\end{aligned}$$

is identically zero for all w. Inverting the feedback $u = \alpha(z) + \beta(z)w$, we obtain $y_T(u, x_0) = y_{GH}(u, \varphi(x_0), z_0)$. □

Example 4.5. Let the systems

$$T = \begin{cases} \dot{x} = u, \\ y_T = x \end{cases} \quad \text{and} \quad H = \begin{cases} \dot{z}_1 = z_2, \\ \dot{z}_2 = u, \\ v = z_1^2, \end{cases}$$

be given. Conditions (i) and (ii) of Theorem 4.4 are clearly verified as well as (iii) is, since $\rho(H) = 1$, $\dim z = 2$, $s_1 = 0$, $s_2 = 1$. In this case there is no need to apply any feedback. By Singh's inversion algorithm we get $u = \psi(z, v, \dot{v}, \ddot{v}) = (\ddot{v} - 2z_2^2)/2z_1$ and $v - z_1^2 = 0$, $\dot{v} - 2z_1z_2 = 0$. Since

$$\frac{\partial}{\partial z}\begin{pmatrix} v - z_1^2 \\ \dot{v} - 2z_1z_2 \end{pmatrix} = \begin{pmatrix} -2z_1 & 0 \\ z_2 & z_1 \end{pmatrix}$$

has rank 2 for $z_1 \neq 0$, we can express z as a function of v, $\dot{v}$ in the neighborhood of any point for which $z_1 \neq 0$. In particular, here we have

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \chi(v, \dot{v}) = \begin{pmatrix} \sqrt{v} \\ \dot{v}/2\sqrt{v} \end{pmatrix} \quad \text{for } z_1 > 0.$$

Then, the compensators

$$G_1 = \begin{cases} \dot{\xi} = \dfrac{\ddot{v} - \dot{v}^2/2v}{2\sqrt{v}} \\ y_G = \xi \end{cases} \quad \text{and} \quad G_2 = \begin{cases} \dot{\xi} = -\dfrac{\ddot{v} - \dot{v}^2/2v}{2\sqrt{v}} \\ y_G = \xi \end{cases},$$

together with the identity map, are local solutions of the G.L.F.P., respectively for $z_1 > 0$ and for $z_1 < 0$.

When a proper compensator is sought, the necessary condition (4.4) must be strengthened into the equality of structures at infinity, and we obtain the following result.

THEOREM 4.6. *The* L.F.P. *is solvable with a proper compensator*

$$G = \begin{cases} \dot{\xi} = f_G(\xi, v), \\ y_G = h_G(\xi, v) \end{cases}$$

only if

$$\rho_i\begin{pmatrix} T \\ H \end{pmatrix} = \rho_i(H) \tag{4.8}$$

for all $i \geqq 1$.

Proof. Let $\mathcal{K}$ denote the field of meromorphic functions in the variables $(x, z, u, \cdots, u^{(N-1)})$, where $N = \dim x + \dim z$. By definition $\rho_i(\begin{smallmatrix} T \\ H \end{smallmatrix}) = \dim \operatorname{span}_{\mathcal{K}} \{dx, dz, d\dot{y}_T, d\dot{v}, \cdots, dy_T^{(i)}, dv^{(i)}\} - \dim \operatorname{span}_{\mathcal{K}} \{dx, dz, d\dot{y}_T, d\dot{v}, \cdots, dy_T^{(i-1)}, dv^{(i-1)}\}$. Denoting by $\mathcal{K}'$ the field of meromorphic functions in the variables $(x, z, \xi, u, \cdots, u^{(N-1)})$, since neither T nor H depend on ξ we have $\rho_i\binom{T}{H} = \dim \operatorname{span}_{\mathscr{K}'}\{dx, dz, d\xi, d\dot{y}_T, d\dot{v}, \cdots, dy_T^{(i)}, dv^{(i)}\} - \dim \operatorname{span}_{\mathscr{K}'}\{dx, dz, d\xi, d\dot{y}_T, d\dot{v}, \cdots, dy_T^{(i-1)}, dv^{(i-1)}\}$. Since $y_{GH} = y_T$, we can substitute $dy_{GH}^{(k)}$ into $dy_T^{(k)}$ for all k in the equation above, thus $\rho_i\binom{T}{H} = \dim \operatorname{span}_{\mathscr{K}'}\{dx, dz, d\xi, d\dot{y}_{GH}, d\dot{v}, \cdots, dy_{GH}^{(i)}, dv^{(i)}\} - \dim \operatorname{span}_{\mathscr{K}'}\{dx, dz, d\xi, d\dot{y}_{GH}, d\dot{v}, \cdots, dy_{GH}^{(i-1)}, dv^{(i-1)}\}$. Moreover, by the properness of (GH), we also have that $dy_{GH}^{(k)} \in \operatorname{span}_{\mathscr{K}'}\{dx, dz, d\xi, d\dot{v}, \cdots, dv^{(k)}\}$, thus $\rho_i\binom{T}{H} = \dim \operatorname{span}_{\mathscr{K}'}\{dx, dz, d\xi, d\dot{v}, \cdots, dv^{(i)}\} - \dim \operatorname{span}_{\mathscr{K}'}\{dx, dz, d\xi, d\dot{v}, \cdots, dv^{(i-1)}\}$. Let $\mathscr{K}''$ denote the field of meromorphic functions in the variables $(z, u, \cdots, u^{(n-1)})$, where $n = \dim z$. Since $dv^{(k)} \in \operatorname{span}_{\mathscr{K}''}\{dz, du, d\dot{u}, \cdots, du^{(k-1)}\}$ for all $k \leqq n$, we finally get

$$\rho_i\binom{T}{H} = \dim \operatorname{span}_{\mathscr{K}''}\{dz, d\dot{v}, \cdots, dv^{(i)}\} - \dim \operatorname{span}_{\mathscr{K}''}\{dz, d\dot{v}, \cdots, dv^{(i-1)}\} = \rho_i(H).$$

5. Conclusion. The model matching and the left factorization problems for nonlinear systems have been considered in the framework of a structural approach based on Singh's inversion algorithm. The result obtained on the model matching problem consists of a computable sufficient condition that also allows us to derive a procedure for constructing a solution, if any.

Necessary computable conditions for the existence of solutions of the left factorization problem have also been found. Under additional technical assumptions, one of the previous conditions has also been shown to be sufficient for assuring the existence of local solutions. However, in this case, the proof is not entirely constructive, since it makes use of the implicit function theorem.

Note added in proof. Further results on the problem have recently been obtained by Huijberts in "A nonregular solution of the nonlinear dynamic disturbance decoupling problem with an application to a complete solution of the nonlinear model matching problem," Memorandum No. 862, University of Twente, 1990.

REFERENCES

[1] G. Conte and A. M. Perdon, *On the causal factorization problem*, IEEE Trans. Automat. Control, 30 (1985), pp. 811-813.

[2] ———, *Zero modules and factorization problems*, Contemp. Math., 47 (1985), pp. 81-94.

[3] J. Descusse and C. H. Moog, *Dynamic decoupling for right invertible nonlinear systems*, Systems Control Lett., 8 (1987), pp. 345-349.

[4] M. D. di Benedetto, *A condition for the solvability of the nonlinear model matching problem*, in Proc. Internat. Conference on Nonlinear Systems, Nantes, France, J. Descusse, M. Fliess, A. Isidori, and D. Leborgne, eds., 1988, Lecture Notes in Control and Information Science, Vol. 122, Springer-Verlag, Berlin, 1989, pp. 102-115.

[5] M. D. di Benedetto, J. W. Grizzle, and C. H. Moog, *A unified notion of rank for a nonlinear system*, in Proc. 27th IEEE Conference on Decision and Control, IEEE Computer Society, Washington, DC, 1988, pp. 926-931.

[6] ———, *Rank invariants for nonlinear systems*, SIAM J. Control Optim., 27 (1989), pp. 658-672.

[7] M. D. di Benedetto and A. Isidori, *The matching of nonlinear models via dynamic state-feedback*, SIAM J. Control Optim., 24 (1986), pp. 1063-1075.

[8] E. Emre, *Generalized model matching and (F, G)-invariant submodules for systems over rings*, Linear Algebra Appl., 50 (1983), pp. 133-166.

[9] E. Emre and M. L. J. Hautus, *A polynomial characterization of (A, B)-invariant and reachability subspaces*, SIAM J. Control Optim., 18 (1980), pp. 420-436.

[10] M. Fliess, *A new approach to the noninteracting control problem in nonlinear systems theory*, in Proc. 23rd Allerton Conference, Monticello, IL, (1985), pp. 123-129.

[11] ———, *A note on the invertibility of nonlinear input-output differential systems*, Systems Control Lett., 8 (1986), pp. 147-151.

[12] J. HAMMER AND M. HEYMANN, *Causal factorization and linear feedback*, SIAM J. Control Optim., 19 (1981), pp. 445-468.

[13] H. J. C. HUIJBERTS, personal communication, 1990.

[14] H. J. C. HUIJBERTS AND H. NIJMEIJER, *Local nonlinear model matching: from linearity to nonlinearity*, Automatica, 26 (1990), to appear.

[15] A. ISIDORI, *Nonlinear Control Systems*, 2nd ed., Communications and Control Engineering Series, Springer-Verlag, Berlin, 1989.

[16] ———, *The matching of a prescribed linear input-output behavior in a nonlinear system*, IEEE Trans. Automat. Control, 30 (1985), pp. 258-265.

[17] 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.

[18] A. ISIDORI AND C. H. MOOG, *On the equivalent of the notion of transmission zeros*, in Proc. IIASA Conf., Sopron, Hungary, C. I. Byrnes and A. Kurszanski, eds., 1986, Lecture Notes in Control and Information Science, Vol. 105, Springer-Verlag, Berlin, 1988, pp. 146-158.

[19] S. KUNG AND I. KAILATH, *Some notes on valuation theory in linear systems*, in Proc. IEEE Conference on Decision and Control, San Diego, 1978, IEEE Computer Society, Washington, DC, 1978, pp. 515-517.

[20] M. MALABRE, *Structure à l'infini des triplets invariants. Application à la poursuite parfaite de modèle*, in Proc. INRIA Conference, Versailles, Lecture Notes in Control and Information Science, Vol. 44, Springer-Verlag, Berlin, 1982, pp. 43-53.

[21] C. H. MOOG, *Nonlinear decoupling and structure at infinity*, Math. Contr. Sign. Syst., 1 (1988), pp. 257-268.

[22] B. C. MOORE AND L. M. SILVERMAN, *Model matching by state feedback and dynamic compensation*, IEEE Trans. Automat. Control, 17 (1972), pp. 491-497.

[23] A. S. MORSE, *Structure and design of linear model following systems*, IEEE Trans. Automat. Control, 18 (1973), pp. 346-354.

[24] H. NIJMEIJER AND W. RESPONDEK, *Decoupling via dynamic compensation for nonlinear control systems*, in Proc. 25th IEEE Conference Decision and Control, Athens, 1986, IEEE Computer Society, Washington, DC, 1986, pp. 192-197.

[25] H. NIJMEIJER AND J. M. SCHUMACHER, *On the inherent integration structure of nonlinear systems*, IMA J. Math. Contr. Inf., 2 (1985), pp. 87-107.

[26] ———, *Zeros at infinity for affine nonlinear control systems*, IEEE Trans. Automat. Control, 30 (1985), pp. 566-573.

[27] T. OKUTAMI AND K. FURUTA, *Model matching of nonlinear systems*, in Proc. IFAC 9th World Congress, J. Gertler and L. Keviczky, eds., 1984, Budapest, Vol. IX, pp. 168-172.

[28] J. F. POMMARET, *Problèmes formels en théorie du contrôle aux dérivées partielles*, C.R. Acad. Sci., Paris Sér. I, 308 (1989), pp. 457-460.

[29] M. K. SAIN AND J. L. MASSEY, *Invertibility of linear time-invariant dynamical systems*, IEEE Trans. Automat. Control, 14 (1969), pp. 141-149.

[30] L. M. SILVERMAN, *Inversion of multivariable linear systems*, IEEE Trans. Automat. Control, 14 (1969), pp. 270-276.

[31] S. N. SINGH, *A modified algorithm for invertibility in nonlinear systems*, IEEE Trans. Automat. Control, 26 (1981), pp. 595-598.

[32] B. F. WYMAN, G. CONTE, AND A. M. PERDON, *Fixed poles in transfer function equations*, SIAM J. Control Optim., 26 (1988), pp. 356-368.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 786-809, July 1991

003

MARKOV DECISION PROBLEMS AND STATE-ACTION FREQUENCIES*

EITAN ALTMAN† AND ADAM SHWARTZ†‡

Abstract. Consider a controlled Markov chain with countable state and action spaces. Basic quantities that determine the values of average cost functionals are identified. Under some regularity conditions, these turn out to be a collection of numbers, one for each state-action pair, describing for each state the relative number of uses of each action. These "conditional frequencies," which are defined *pathwise*, are shown to determine the "state-action frequencies" that, in the finite case, are known to determine the costs. This is extended to the countable case, allowing for unbounded costs. The space of frequencies is shown to be compact and convex, and the extreme points are identified with stationary deterministic policies.

Conditions under which the search for optimality in several optimization problems may be restricted to stationary policies are given. These problems include the standard Markov decision process, as well as constrained optimization (both in terms of average cost functionals) and variability-sensitive optimization. An application to a queueing problem is given, where these results imply the existence and explicit computation of optimal policies in constrained optimization problems.

The pathwise definition of the conditional frequencies implies that their values can be controlled directly; moreover, they depend only on the limiting behavior of the control. This has immediate application to adaptive control of Markov chains, including adaptive control under constraints.

Key words. Markov decision process, average cost, constrained optimization, state-action frequencies, nonstationary control

AMS(MOS) subject classifications. 90B50, 60G17, 60J10, 93E20

Introduction. An important issue in optimization problems is the reduction of the space of policies over which we optimize. This is motivated by the need to reduce the complexity of the search for optimal policies, and by the desire to restrict attention to those policies that are easy to implement. Indeed, in many optimization problems we first show that it is possible to restrict the optimization to the class of stationary policies. This simplifies the search, since many computational methods are available in the stationary case. Furthermore, the implementation requires relatively little memory. Conditions that ensure that we may indeed restrict the search of optimal policies for Markov decision processes to stationary policies (or even to deterministic stationary policies) are an active area of research (see, e.g., Borkar [9], [10], Cavazos-Cadena [12], Dekker and Hordijk [14], Sennott [28], [29] and references therein).

On the other hand, it is of interest to know how flexible we can be in the choice of policies, in a way that does not change the values of average cost criteria. This is the case in adaptive optimization, where we often use on-line estimation schemes to generate an approximation of the optimal control (the certainty equivalence approach). The goal in this case is to achieve the same performance as in the case of full information.

These two issues are treated in this paper in the framework of the following question. For a given policy, what are the quantities that determine the values of average cost functionals? Fix a state x and an action a. For each t, consider the (random) number of times the process visited state x and action a was used by time t. It turns out that in many cases average costs are determined by the limits (in time) of the expectations of such "state-action frequencies." For each time t, consider the (random) ratio of the number of uses of action a while in state x, to the number of

* Received by the editors November 14, 1988; accepted for publication (in revised form) June 15, 1990.

† Electrical Engineering, Technion—Israel Institute of Technology, Haifa, 32000 Israel.

‡ The research of this author was performed in part while he was visiting the Mathematical Sciences Research Center, AT&T Bell Laboratories.

visits to state x. Below we show that the pathwise limits of these "conditional frequencies" are the more basic quantity, in that they determine the expected state action frequencies.

We deal with countable state and action spaces, and obtain classes of policies that achieve every possible state action frequency; we term such classes "complete." In the finite case, some questions of completeness are investigated in [2], [15], and [22]; Derman [15] gives conditions for the completeness of Markov policies. Hordijk and Kallenberg [22] strengthen this result to Markov policies having just one accumulation point of the "matrix" of frequencies. Derman [15] and, later, Hordijk and Kallenberg [22] give conditions for the completeness of stationary policies. Two time sharing policies were introduced by Altman and Shwartz [1]-[3], who show that under the conditions of Derman [15], completeness depends on pathwise limit properties only, and in particular may be achieved using deterministic (but nonstationary) policies. In this paper we show that in the countable case the space of achievable frequencies is a compact convex set whose extreme points are frequencies obtained by deterministic (stationary) policies. This extends the geometric characterization given in the finite case by Derman [15], Hordijk and Kallenberg [22], and Altman and Shwartz [2].

We give conditions under which some classes of policies (such as the stationaries) are "sufficient" in the countable case for several optimization problems, including optimization under several constraints. These results allow the use of steady-state analysis of systems, which simplifies the search for optimal policies considerably. It becomes possible to translate results on performance, which in many cases deal with "steady state," into results concerning optimization (see, e.g., Altman and Shwartz [1], [3]). Previous results on the sufficiency of stationary policies in the case of countable state space dealt only with the minimization (or maximization) of a single criterion.

Then, we introduce a larger family of "sufficient" policies—the action time sharing (*ats*) policies—which is characterized by the existence of a with probability one limit to the conditional frequencies. In contrast with the standard "small" classes of policies such as the stationary policies, these policies are flexible enough to be useful for adaptive problems, as they have the following important property: the expected frequencies (and thus the cost) achieved by any policy depends only on the (pathwise) limiting behavior of the control mechanism. More precisely, it depends only on the limit of the conditional frequencies, described above. Therefore it is possible to use nonstationary algorithms based on real-time estimation of unknown parameters, and still obtain optimality. Moreover, whereas existing results on adaptive control of Markov chains consider only the optimization of a single criterion, the present results can be used to obtain adaptive controls under more general criteria, such as constrained optimization. An application of these ideas in the case of *finite* state and action spaces is given in Altman and Shwartz [2, § 5], [4]. The computation of optimal policies of the *ats* type is equivalent to the computation over the more restricted class of stationary policies, and the implementation is just as simple.

After introducing the model and some notation, § 1 provides the basic motivation by introducing the standard Markov decision problem and a constrained optimization problem. In § 2 we derive conditions under which the frequencies determine the value of optimization criteria, and under which stationary policies or other complete classes of policies are sufficient for the two optimization problems. In § 3 the basic results concerning the completeness of the stationary policies and the role of the conditional frequencies in determining the behavior of the process are derived. Since the state and action spaces are countable (and thus not compact), a tightness condition is used. The literature concerning the tightness problem is extensive; in § 4 we adapt some applicable results. The case where tightness does not hold is treated by imposing conditions on the cost, under which tight policies are "better" than nontight ones. In § 5 it is shown that the space of frequencies is compact, and has the geometric characterization as the convex hull of the frequencies of stationary deterministic policies. This has implications to the existence of optimal policies in constrained optimization problems. Finally, we apply and extend the results of the previous sections. In § 6 we treat a queueing network, and in § 7 an equivalence between the constrained optimization problem and an associated linear program (which is well known in the finite case [15], [22]) is extended to the countable case. Section 8 treats some lesser known optimization problems involving variance.

1. The model and the problems. Let $\{X_t\}_{t=1}^{\infty}$ be a discrete time process defined on the countable *state space* $\mathbf{X}=\{0, 1, \cdots\}$. At time t an action A_t from the countable *action space* $\mathbf{A}$ is taken. Denote by $A(x)$ the set of actions available when in state x. $h_n := (X_1, A_1, X_2, A_2, \cdots, X_n, A_n)$ is the *history* of $\{X_t\}$. Denote the transition probabilities for the controlled Markov chain by

$$(1.1)\quad P_{xay} := P(X_{n+1}=y\,|\,X_n=x;\ A_n=a)=P(X_{n+1}=y\,|\,h_{n-1}=h,\ X_n=x;\ A_n=a).$$

A policy u in the *policy space* U is defined by $u=\{u_1, u_2, \cdots\}$, where u_t is applied at time epoch t, and $u_t(\cdot\,|\,h_{t-1}, X_t)$ is a conditional probability measure over $\mathbf{A}$. Each policy u induces a probability measure P^u on the space of paths (which serves as the canonical sample space Ω). The corresponding expectation operator is denoted by E_u.

A *Markov policy* $f\in U(M)$ is characterized by the dependence of $u_t(\cdot\,|\,h_{t-1}, X_t)$ on X_t only; i.e., $u_t(\cdot\,|\,h_{t-1}, X_t)=u_t(\cdot\,|\,X_t)$. A *stationary policy* $g\in U(S)$ is characterized by a single conditional probability measure $u(\cdot\,|\,X_t)=p^g_{\cdot|X_t}$ over $\mathbf{A}$, so that $p^g_{A(x)|x}=1$; under g, $\{X_t\}$ becomes a Markov chain with stationary transition probabilities, given by $P^g_{xy}=\sum_{a\in A(x)} p^g_{a|x}P_{xay}$. The class of stationary deterministic policies $U(SD)$ is a subclass of $U(S)$, and every $g\in U(SD)$ is identified with a mapping $g:\mathbf{X}\to\mathbf{A}$, so that for each x, $p^g_{\cdot|x}=\delta_{g(x)}(\cdot)$ is concentrated at one point $a\in A(x)$.

Let $c(x, a)$ be a real valued function on $\mathbf{X}\times\mathbf{A}$, possibly unbounded, and let

$$C^t_x(u)=\frac{1}{t}E_u\left[\sum_{s=1}^{t} c(X_s, A_s)\,|\,X_1=x\right].$$

We assume throughout that for each u, the cost $C^t_x(u)$ is well defined. This will usually follow from uniform integrability assumptions on $c(X_t, A_t)$, or from a condition that $c(\cdot,\cdot)$ is bounded below. The optimization problem OP involves the minimization of average cost functionals:

$$(1.2\text{a})\qquad \bar{C}_x(u)=\overline{\lim_{t\to\infty}}\, C^t_x(u),$$

$$(1.2\text{b})\qquad \underline{C}_x(u)=\underline{\lim}_{t\to\infty}\, C^t_x(u).$$

These include the standard "positive" and "negative" Markov decision problems. Given the constants V_k, $1\leqq k\leqq K$, the constrained optimization problem COP is defined as

$$(1.3\text{a})\qquad \text{minimize}\quad \bar{C}_x(u)\qquad \text{subject to}\quad \bar{D}^k_x(u)\leqq V_k,\quad 1\leqq k\leqq K,$$

$$(1.3\text{b})\qquad \text{minimize}\quad \underline{C}_x(u)\qquad \text{subject to}\quad \bar{D}^k_x(u)\leqq V_k,\quad 1\leqq k\leqq K,$$

where $\bar{D}^k_x(u)$ is defined similarly to $\bar{C}_x(u)$ with $c(x, a)$ replaced by $d^k(x, a)$, and both $c(x, a)$ and $d^k(x, a)$ may be unbounded. For finite state and action spaces, a solution to the constrained optimization problem based on linear programming was already obtained by Derman [15] and Hordijk and Kallenberg [22], and some variables of this linear program are limits of the state-action frequencies (1.4).

These *expected state-action frequencies* (Derman [15]) and *expected state frequencies*

$$\text{(1.4)}\qquad \begin{aligned} \bar f^T_{x,u}(y,a) &:= \frac{1}{T}\sum_{s=1}^{T} P^u(X_s=y, A_s=a \mid X_1=x), \\ \bar f^t_{x,u}(y) &:= \frac{1}{t}\sum_{s=1}^{t} P^u(X_s=y \mid X_1=x) \end{aligned}$$

are key quantities in the analysis below. Let the "matrix" $\{\bar f_{x,u}(y,a)\}_{y,a}$ denote a generic accumulation point of the infinite "matrix" $\{\bar f^T_{x,u}(y,a)\}_{y,a}$ as $T\to\infty$ (i.e., an accumulation point in a countable-dimensional space with one coordinate for each state action pair), and let $\{\bar f_{x,u}(y)\}_y$ denote any accumulation point of the infinite vector $\{\bar f^t_{x,u}(y)\}_y$. Let $\bar F_{x,u}$ denote the set of all limit matrices $\bar f_{x,u}(\cdot,\cdot)$. Any class of policies U' determines a set of accumulation points $L_x(U') := \bigcup_{u\in U'} \bar F_{x,u}$ and the set of all such limits is denoted by $L_x := \bigcup_{u\in U} \bar F_{x,u}$. We use the abbreviations $L_x(S) := L_x(U(S))$ and $L_x(D) := L_x(U(SD))$.

The following definitions are useful for the sample-path analysis of § 3. Let $f^T(y,a) := (1/T)\sum_{s=1}^T 1\{X_s=y, A_s=a\}$ denote the sample-path frequency at which the event of being at state y and choosing action a occurs till time T. The expectation of the random variable $f^T(y,a)$ under u starting at x is thus $\bar f^T_{x,u}(y,a)$. The frequency at which the event of being at state y occurs till time T is denoted by $f^T(y) = (1/T)\sum_{s=1}^T 1\{X_s=y\}$. Finally, $f^T(a\mid y) := f^T(y,a)[f^T(y)]^{-1}$ is the frequency at which action a is chosen conditioned on being in state y, until time T. If $f^T(y)=0$ define $f^T(a\mid y) := 0$. Denote by $f(y,a)$ (respectively, $f(y)$) any accumulation point of $f^T(y,a)$ (respectively, $f^T(y)$) as $T\to\infty$.

Let g be a stationary policy. The following standard result will be frequently used.

LEMMA 1.1. *Assume that under g the process $\{X_t\}_{t=1}^\infty$ has one positive recurrent class, and that from any transient state, absorption into the recurrent class occurs in finite expected time. Then*

$$\bar f_{x,g}(y,a) = \pi^g(y)p^g_{a|y} = \lim_{t\to\infty} f^t(y,a) \quad P^g \textit{ a.s.}$$

For the last equality to hold, it suffices that absorption occurs with probability one.

A class of policies U' is called *complete* if $L_x = \bigcup\{\bar F_{x,u'}\colon u'\in U'$ and $\bar F_{x,u'}$ *is a singleton*$\}$. U' is called *weakly complete* if

$$L_x \cap \left\{\zeta\colon \sum_{y,a}\zeta(y,a)=1\right\} \subset \bigcup\{\bar F_{x,u'}\colon u'\in U' \text{ and } \bar F_{x,u'} \textit{ is a singleton}\}.$$

Note that for each t, $\bar f^t_{x,u}(y,a)$ can be considered as a probability measure over $\mathbf{X}\times\mathbf{A}$. The condition $\sum_{y,a}\bar f_{x,u}(y,a)=1$ for every limit point $\bar f_{x,u}(y,a)$ of a subsequence $\{\bar f^{t_n}_{x,u}(y,a)\}_n$ is equivalent to tightness of this set of measures [8]. Thus weak completeness considers only tight frequencies.

A class of policies U' is called *sufficient* for an optimization problem if for any policy u there is a policy $u'\in U'$ that performs at least as well. The motivation for studying questions of completeness and the spaces of frequencies is provided in § 2 below, where the connection between completeness and sufficiency is established. Note that sufficiency does not imply existence of an optimal policy, but rather that the search for "good" policies can be restricted to any subclass that is sufficient.

The following assumptions are used frequently in the paper:

(A0) At each state x, the set of available actions $A(x)$ is finite.

(A1) Under any policy $g \in U(S)$ the state space contains a single *positive* recurrent class, and absorption into the positive recurrent class takes place in finite expected time.

It follows from Fisher [18] that under (A0) and (A1), if there are no transient states under any policy in $U(SD)$ then the chain is ergodic under each policy in $U(S)$ if and only if it is ergodic under each policy in $U(SD)$ (see also § 5, Corollary 5.3).

(A2(u)) Given a policy u, the expected frequencies $\{\bar{f}^{t}_{x,u}(y, a)\}_t$ are tight.

(A2) Assumption (A2(u)) holds for all policies $u \in U$.

(A2*) The family of stationary probabilities corresponding to policies in $U(SD)$ is tight.

Remarks. (i) The issue of tightness is treated in § 4. In Lemma 4.1 we show that under the appropriate conditions, (A2) is equivalent to (A2*). We give simple verifiable sufficient conditions for (A2*), and develop some methods for the nontight case.

(ii) Assumption (A2(u)) depends on the initial state x, even when (A1) holds. For example, let u' be a policy that violates (A2(u')) (e.g., the policy constructed in [17]). Let g be a policy for which (A2(g)) holds (under (A1), this holds for any stationary policy). If u equals u' whenever $X_0 = x$ and otherwise uses g, then clearly (A2(u)) holds for all initial states except x. Throughout the paper, reference to (A2(u)) will implicitly assume a fixed initial state, which is omitted from the notation.

To make the discussion more concrete, we cite Theorem 3.2, whose proof is given in § 3.

THEOREM 3.2. *Under* (A1) *the class of stationary policies is weakly complete.*

As will become clear in § 3, the property of completeness does not depend on stationarity; it is more naturally defined through conditional frequencies. This will be seen to provide a large degree of flexibility, which can be applied in a straightforward manner to adaptive optimization problems [4].

2. Sufficiency and completeness. The aim of this section is to establish the relation between optimization problems and state-action frequencies, and in particular between sufficiency and completeness. In the case of finite state and action spaces it is known that the time average expected cost has a representation as a linear function of the expected state-action frequencies (e.g., [15]). We extend this result to the countable case, and establish sufficient conditions under which the costs (1.2a) and (1.2b) can be represented as linear functionals (2.4) of the frequencies. The advantage of this approach is that it deals directly with the cost functionals, and therefore applies to many classes of optimization problems. In the following sections we investigate the optimization problems OP and COP, and show the connection between completeness and sufficiency. In particular, we present conditions under which the search for solutions of OP and COP can be restricted to those policies for which the costs have the linear representation (2.4) in terms of the frequencies. Similar results are obtained in § 8 for other optimization problems. These results motivate the further investigation of the achievable frequencies under various classes of policies, which is carried out in § 5. We will be especially interested in finding out which classes of policies are complete. This will indicate when a class of policies is sufficient for the optimization problems OP and COP, or, in other words, whether we may restrict the search for optimal policies to smaller classes of policies. Moreover, as will become clear is § 3, this approach identifies the key quantities that determine the costs, and allows for a flexible choice of controls while keeping the cost fixed.

The results of this section concerning optimization problems are given under condition (A2), which is a rather strong "uniform stability" assumption. In § 4 we provide natural sufficient conditions for (A2), and also show how the results can be extended when tightness does not hold.

2.1. Representation of costs through frequencies. Unlike the case of finite state and action spaces, the time average expected cost in the countable case need not even have a representation as a function of the expected state-action frequencies.

Counterexample 2.1. The deterministic case. Let $P_{xy} = 1\{x+1=y\}$ and $c(x, a) = 1$ for all states and actions. The action is thus irrelevant to both the dynamics and cost and we may assume that there is just one possible action. Under any policy u we clearly have $\bar{f}_{x,u}(y, a) = 0$, while $C_x(u) = 1$.

In this example (A1) does not hold. Counterexample 3.5 in § 3 presents a case where (A1) holds but (A2) does not, and which exhibits similar behavior.

Lemmas 2.2 and 2.3 provide conditions under which a linear representation (2.4) holds. Fix an initial state x and a policy u. Since by assumption $C_x^t(u)$ is well defined, the definitions imply

$$\bar{C}_x(u) = \overline{\lim_{t\to\infty}} \sum_{y,a} \bar{f}^t_{x,u}(y, a) c(y, a). \tag{2.1}$$

Let $\{s_n\}_n$ be a subsequence along which the limit is obtained, i.e.,

$$\bar{C}_x(u) = \lim_{n\to\infty} \sum_{y,a} \bar{f}^{s_n}_{x,u}(y, a) c(y, a). \tag{2.2}$$

Using diagonalization, choose a further subsequence $\{t_n\}_n$ so that $\bar{f}^{t_n}_{x,u}(y, a) \to \bar{f}_{x,u}(y, a)$, for all y and a.

LEMMA 2.2. *Assume* (A1) *and let* $\{c(X_s, A_s)\}_s$ *be uniformly integrable under* P^u. *If* (A2(u)) *holds then the costs* (1.2) *are a function* (2.4) *of the* $\bar{f}_{x,u}$ *defined above. If* v *is a policy such that* $\bar{F}_{x,v} = \{\bar{f}_{x,u}\}$ *and* $\{c(X_s, A_s)\}_s$ *are also uniformly integrable under* P^v, *then* $\bar{C}_x(u) = \bar{C}_x(v)$ *and* $\underline{C}_x(u) = \underline{C}_x(v)$.

Proof. Consider first the cost function defined through (1.2a). Note that for each t, $\bar{f}^t_{x,u}(\cdot, \cdot)$ can be considered as a probability measure over $\mathbf{X} \times \mathbf{A}$, and the cost $c(\cdot, \cdot)$ can then be viewed as a random variable over $\mathbf{X} \times \mathbf{A}$. The convergence $\bar{f}^{t_n}_{x,u}(y, a) \to \bar{f}_{x,u}(y, a)$ for all y and a thus translates under (A2(u)) into weak convergence of probability measures along t_n. As $\{c(X_s, A_s)\}_s$ is uniformly integrable with respect to P^u, $c(\cdot, \cdot)$ is also uniformly integrable with respect to $\{\bar{f}^t_{x,u}\}_t$; this follows from the fact that for every function h,

$$\sum_{y,a} \bar{f}^t_{x,u}(y, a) h[c(y, a)] = E_u \frac{1}{t} \sum_{s=1}^{t} [h[c(X_s, A_s)] \mid X_1 = x]. \tag{2.3}$$

This weak convergence of $\bar{f}^t_{x,u}$ implies the convergence of $c(\cdot, \cdot)$ in distribution, and combining this to the uniform integrability of $c(\cdot, \cdot)$ we finally obtain [8]

$$\bar{C}_x(u) = \sum_{y,a} \bar{f}_{x,u}(y, a) c(y, a). \tag{2.4}$$

The argument for (1.2b) is identical. The last claim is now immediate since $\bar{f}_{x,u} = \bar{f}_{x,v}$. □

It is not difficult to establish the following converse to Lemma 2.2. If (2.4) holds for each limit point $\bar{f}_{x,u}$ in $\bar{F}_{x,u}$ and c satisfies $\infty > \bar{c} > c(y, a) > \varepsilon > 0$, for all y, a, then (A2(u)) holds. But for an arbitrary c (2.4) may not imply (A2(u)) (for example, $c = 0$ provides no information).

Next, we discuss the representation (2.4) for stationary policies. Fix $g \in U(S)$ and an initial state x, and let $0 \in \mathbf{X}$ be recurrent under g. With the standard convention that $\inf \varnothing := \infty$, define $\eta(1) := \inf\{t \geqq 1: X_t = 0\}$, $\eta(k+1) := \inf\{t > \eta(k): X_t = 0\}$, where $\eta(k) = \infty$ implies $\eta(k+1) = \infty$.

LEMMA 2.3. *Assume* (A1) *and let* $g \in U(S)$. *Then* (A2(g)) *holds, and the representation* (2.4) *for the costs* (1.2) *holds whenever one of the following is true*: (i) $\{c(X_s, A_s)\}_s$ *are uniformly integrable with respect to* P^g; (ii) *c is bounded from below and* (A3(g)) *holds*; (iii) *c is bounded from above and* (A3(g)) *holds, where*

$$\text{(A3}(g)\text{)} \qquad \begin{aligned} &E_g\left[\sum_{s=1}^{\eta(1)-1} |c(X_s, A_s)| \,|X_1 = x\right] = \infty \\ &\textit{implies } E_g\left[\sum_{\eta(1)}^{\eta(2)-1} |c(X_s, A_s)| \,|X_1 = x\right] = \infty. \end{aligned}$$

Note that if, under g, x belongs to the recurrent class then (A3(g)) always holds (see the proof below). In particular, when there are no transient states under g, (A3(g)) holds.

Proof. The first claim follows from Lemma 1.1, and (i) is obtained by combining this with Lemma 2.2.

To prove (ii), consider first a cost of the form $c(x, a) = c(x)$. Recall that the initial condition is fixed, and is omitted from the notation below. Denote by $\tau_k := E_g[\eta(k+1) - \eta(k)]$ the expected time between consecutive visits to state zero under g. (We call such a period a cycle.) From Chung [13], under (A1), $\tau := \tau_k$ is independent of k. Denote the (sample) cost over the kth cycle by $Y(k) := \sum_{t=\eta(k)}^{\eta(k+1)-1} c(X_t)$. If $\eta(k) = \infty$ set $Y(k) := 0$. Assume that

$$(*) \qquad E_g\left[\sum_{\eta(1)}^{\eta(2)-1} |c(X_s)|\right] < \infty.$$

Denote by $W := E_g[Y(1)]$ the total expected cost during the first cycle. Since c is bounded below, (A1) implies that (*) is equivalent to W being finite. From Chung [13] it follows that under (A1), (*), and (A3(g)),

$$(2.5) \qquad \bar{C}_x(g) = \underline{C}_x(g) = C(g) = \frac{W}{\tau} = \sum_{y \in \mathbf{X}} \pi_y^g c(y).$$

But as g is stationary, $\bar{f}_{x,g}(y) = \pi_y^g$ whereas the tightness impliess $\sum_{a \in \mathbf{A}} \bar{f}_{x,g}(y, a) = \bar{f}_{x,g}(y)$. Hence

$$(2.6) \qquad C(g) = \sum_{y \in \mathbf{X}} \bar{f}_{x,g}(y) c(y) = \sum_{y,a} \bar{f}_{x,u}(y, a) c(y).$$

Next, if (*) does not hold then $W = \infty$. Denote $c^M(x) := c(x) 1\{c(x) \leqq M\}$, and define W^M, $Y^M(k)$, and $C_x^M(u)$ as before, but with $c^M(x)$ replacing $c(x)$. The following monotone convergence holds pathwise:

$$Y(1) = \sum_{s=\eta(1)}^{\eta(2)-1} c(X_s) = \lim_{M \to \infty} \sum_{s=\eta(1)}^{\eta(2)-1} c^M(X_s) = \lim_{M \to \infty} Y^M(1).$$

Clearly, $C(g) \geqq \overline{\lim}_{M \to \infty} C_x^M(g) = \overline{\lim}_{M \to \infty} W^M/\tau = W/\tau$ by (2.5) and the monotone convergence theorem. Thus $C(g) = W/\tau$, i.e., all but the last equality in (2.5) hold independently of (*). It then follows from (2.5) that

$$(2.7) \qquad \infty = \lim_{M \to \infty} \frac{W^M}{\tau} = \lim_{M \to \infty} \sum_{y \in \mathbf{X}} \pi_y^g c^M(y) = \sum_{y \in \mathbf{X}} \pi_y^g c(y)$$

using the monotone convergence theorem. The argument leading to (2.6) now implies (2.4).

Finally, we allow the cost to depend on the action. Let $\hat{c}(x) := \sum_{a\in A(x)} p^g_{a|x} c(a, x)$. Note that

$$C_x(g) = \overline{\lim_{t\to\infty}} \frac{1}{t} E_u\left[\sum_{s=1}^{t} \hat{c}(X_s) \,|\, X_1 = x\right] = \sum_{y\in\mathbf{X}} \bar{f}_{x,g}(y)\hat{c}(y)$$

$$= \sum_{y\in\mathbf{X}} \sum_{a\in\mathbf{A}} \bar{f}_{x,g}(y) p^g_{a|y} c(y, a) = \sum_{y,a} \bar{f}_{x,g}(y, a) c(y, a),$$

since $\bar{f}_{x,g}(y) p^g_{a|y} = \bar{f}_{x,g}(y, a)$, and where changes of order of summation are justified since c is bounded below. The proof in the case the cost bounded from above is identical. □

That (A3(g)) is necessary can be seen through the following example. Let $\mathbf{X}$ be the positive integers and let $\mathbf{A} := \{a\}$. Let $P_{0a0} = 1$, and $P_{xa0} = 0.5$, $P_{xa(x+1)} = 0.5$ for $x > 0$. With $c(y, a) = b^y$, (A3(g)) is violated and (2.4) fails to hold whenever $b \geqq 2$.

2.2. Optimization and frequencies.

Using the previous lemmas we next discuss optimization under the expected average cost criterion.

LEMMA 2.4. *Assume $c(x, a)$ is bounded below, and let $C_x(u)$ denote either of the costs* (1.2). *If*

$$C_x(u) = \lim_{n\to\infty} \sum_{y,a} \bar{f}^{t_n}_{x,u}(y, a) c(y, a) \tag{2.8}$$

for some u and sequence $\{t_n\}$, then for any accumulation point $\bar{f}_{x,u}$ of $\{\bar{f}^{t_n}_{x,u}\}$, $C_x(u) \geqq \sum_{y,a} \bar{f}_{x,u}(y, a) c(y, a)$.

Proof. Assume first that c is positive and let $\{s_n\}$ be a subsequence of $\{t_n\}$ such that $\bar{f}^{s_n}_{x,u} \to \bar{f}_{x,u}$ for all (a, y). An application of Fatou's lemma, where c is considered as a σ-finite measure yields

$$C_x(u) = \lim_{n\to\infty} \sum_{y,a} \bar{f}^{s_n}_{x,u}(y, a) c(y, a) \geqq \sum_{y,a} \underline{\lim}_{n\to\infty} \bar{f}^{s_n}_{x,u}(y, a) c(y, a) = \sum_{y,a} \bar{f}_{x,u}(y, a) c(y, a). \tag{2.9}$$

In general, shifting c to obtain such a measure, the same argument applies. □

COROLLARY 2.5. *Let u and v be two policies such that $\bar{F}_{x,v} = \{\bar{f}_{x,v}\}$, and $\bar{f}_{x,u} = \bar{f}_{x,v}$, for some accumulation point $\bar{f}_{x,u} \in \bar{F}_{x,u}$. Assume that under v the representation* (2.4) *holds and $c(x, a)$ is bounded below. Then $C_x(v) \leqq C_x(u)$, where $C_x(u)$ stands for either of the costs* (1.2).

The following theorem gives conditions under which the search for optimal (or ε optimal) policies can be restricted to a subclass of policies.

THEOREM 2.6. *Consider the problem of minimizing $\bar{C}_x(u)$ (or minimizing $\underline{C}_x(u)$). Assume* (A1) and (A2) *and let U' be complete. Then U' is sufficient if one of the following assumptions holds*:

(i) *$\{c(X_s, A_s)\}_s$ is uniformly integrable with respect to P^u for each $u \in U$.*

(ii) *For each $u' \in U'$* (2.4) *holds and $c(x, a)$ is bounded below.*

Proof. (i). For any $u \in U$, there exists a $v \in U'$ satisfying the hypotheses of Lemma 2.2, and sufficiency follows. The proof of (ii) follows from Lemma 2.4 and Corollary 2.5. □

Note that the question of existence of optimal policies is not raised here.

2.3. Constrained optimization.

The reason for restricting problem COP to cost functionals D^k defined through (1.2a) is that, when the constraints are defined through (1.2b), a complete class of policies may not be sufficient, even if the state and action spaces are finite. For example:

Counterexample 2.7. Optimization under a constraint. Let $\mathbf{X}=\{x\}$, and $\mathbf{A}=\{p, q\}$. Set $c(x,p)=d(x,q)=-1$, $d(x,p)=c(x,q)=0$. Define $\underline{C}$ and $\underline{D}$ through (1.2b). The objective is to minimize $\underline{C}_x(u)$ under the constraint $\underline{D}_x(u)\leqq -0.5$.

In the finite, single class case, it is well known that the class of stationary policies achieves all possible frequencies. It is easy to see that the best stationary policy g chooses p or q with equal probability, and $\underline{C}_x(g)=-0.5=\underline{D}_x(g)$. Consider the policy u that uses p at times $(2n)^2\leqq t<(2n+1)^2$ $n=1,2,\cdots$ and action q at the remaining epochs. Then $\underline{C}_x(u)=\underline{D}_x(u)=-1$, and we conclude that there is no stationary optimal policy.

THEOREM 2.8. *Consider problem* COP (1.3a) *and* (1.3b). *Under* (A1) *and* (A2) *the stationary policies are sufficient if one of the following holds*;

(i) $\{c(X_s, A_s)\}_s$ *and* $\{d^k(X_s, A_s)\}_s$, $1\leqq k\leqq K$ *are uniformly integrable with respect to* P^u *for each u, or*

(ii) $c(x, a)$ *and* $d^k(x, a)$, $1\leqq k\leqq K$ *are bounded below and* (A3(g)) *holds for all* $g\in U(S)$, *with respect to c and* d^k, $1\leqq k\leqq K$.

Remark. It clearly suffices to check (A3(g)) for those policies that satisfy the constraints (see also § 4).

Proof. Consider first (1.3a). Fix an arbitrary policy u. Let t_n be a subsequence such that $\bar{C}_x(u)=\lim_{n\to\infty}\sum_{y,a}\bar{f}^{t_n}_{x,u}(y,a)c(y,a)$, and such that the limits $\bar{f}^{t_n}_{x,u}(y,a)\to \bar{f}_{x,u}(y,a)$ for all y and a, and $\lim_{n\to\infty}\sum_{y,a}\bar{f}^{t_n}_{x,u}(y,a)d^k(y,a)$ exist. According to Corollary 3.3 the class of stationary policies $U(S)$ is complete, hence there exists a stationary policy g such that $\bar{f}_{x,g}=\bar{f}_{x,u}$. Under assumption (i), Lemma 2.3 implies that (A2(g)) holds, and so Lemma 2.2 implies $\bar{C}_x(g)=\bar{C}_x(u)$. Finally,

$$\bar{D}^k_x(u)=\overline{\lim_{t\to\infty}}\sum_{y,a}\bar{f}^{t}_{x,u}(y,a)d^k(y,a)\geqq\overline{\lim_{n\to\infty}}\sum_{y,a}\bar{f}^{t_n}_{x,u}(y,a)d^k(y,a)$$

$$=\sum_{y,a}\bar{f}_{x,u}(y,a)d^k(y,a)=\bar{D}^k_x(g),$$

where the next to last equality is validated by the arguments of the proof of Lemma 2.2. This proves the theorem under assumption (i).

Under (ii), since (A1) implies that $\bar{f}_{x,g}$ is a singleton, Lemma 2.3 and Corollary 2.5 can be invoked to conclude $\bar{C}_x(g)\leqq\bar{C}_x(u)$ and $\bar{D}^k_x(g)\leqq\bar{D}^k_x(u)$ for each k, and the proof for (1.3a) is concluded. The proof for (1.3b) is identical. □

Finally, we consider an arbitrary complete class.

COROLLARY 2.9. *Assume* (A1) *and* (A2) *and consider* COP (1.3). *Let* U' *be any complete class of policies. Assume* (2.4) *holds for all* $u'\in U'$ *and for c and* d^k. *Then* U' *is sufficient if one of the following holds*;

(i) $\{c(X_s, A_s)\}_s$ *and* $\{d^k(X_s, A_s)\}_s$, $1\leqq k\leqq K$ *are uniformly integrable with respect to* P^u *for each u, or*

(ii) $C(x, a)$ *and* $d^k(x, a)$, $1\leqq k\leqq K$ *are bounded below.*

Proof. Stationarity is used in the proof of Theorem 2.8 solely to guarantee that $\bar{F}$ is a singleton and the representation (2.4) holds. □

3. Completeness: action time sharing. In Theorem 3.2 we prove that the class of stationary policies is weakly complete under (A1), and complete (Corollary 3.3) under the additional assumption (A2). This and the results of § 2 allow us to recover and extend classical results, concerning optimality of stationary policies.

The classical approach to Markov optimization problems relies on the specific class of stationary policies, and on their statistical properties. In contrast, the point of view taken here is to find weak sufficient conditions for a class of policies to be complete. The class of "action time sharing" policies introduced below includes the stationary policies. However, the novelty of this approach is expressed in Theorem 3.6, which states that the frequencies achieved by "*ats*" policies depend only on their **pathwise** conditional frequencies. This implies (Theorem 3.7) that completeness can be achieved within subclasses other than stationaries: for example, using deterministic, nonstationary policies.

Fix $\alpha := \{\alpha_y^a, y \in \mathbf{X}, a \in A(y)\}$, where $\alpha_y^a \geqq 0$, and $\sum_{a \in A(y)} \alpha_y^a = 1$ for each $y \in \mathbf{X}$.

DEFINITION. A policy u is an "action time sharing" (*ats*) policy with parameter α, and is denoted as $\hat{\alpha}$ if for every state y that is visited infinitely often P^u almost surely and any action a,

$$f^T(a \mid y) \to \alpha_y^a \quad \text{as } T \to \infty \quad P^u \text{ a.s.}$$

Thus an *ats* policy $\hat{\alpha}$ alternates between several actions at each state so as to achieve a prescribed (state dependent) limiting relative frequency for each action. There are no restrictions as to how the limiting frequencies $f^T(a \mid y)$, are achieved, and there are many ways such a policy can be realized.

A realization of an *ats* policy with parameter α is a mapping h from the space S_α of all possible collections α to the space of all policies U. Given such a mapping h, let $U^h(ats) := h(S_\alpha)$ denote the subclass of *ats* policies of the form $\hat{\alpha} = h(\alpha)$ for some $\alpha \in S_\alpha$. For example, setting $p_{a|y} = \alpha_y^a$ defines a stationary policy, where $p_{a|y}$ are the conditional distributions. We thus recover the class of stationary policies, where the realization is by randomly choosing the actions using unfair dice. Another possible realization of *ats* policies is through the use of a counter for each $y \in \mathbf{X}$, $a \in A(y)$. We then choose in a *deterministic way* which action to use for every state, so that the appropriate conditional frequencies are achieved.

The main result of this section, Theorem 3.7, states that under (A1) and (A2), $U^h(ats)$ is complete for any h (see the definition of completeness in § 1). Moreover, the frequencies $\bar{f}_{x,\hat{\alpha}}(y, a)$ depend only on the parameter α, and not on the realization $\hat{\alpha} = h(\alpha)$ (Theorem 3.6). We proceed to investigate the completeness of stationary policies, and will then turn to *ats* policies. But first we need a technical lemma.

LEMMA 3.1. *Under* (A1) *for any transition matrix* P^g, $g \in U(S)$ *there exists a measure* π *such that*

$$\pi(y) \geqq \sum_{z \in \mathbf{X}} \pi(z)[P^g]_{zy}. \tag{3.1}$$

The measure π *is finite, is unique up to a multiplicative constant, and in fact* $\pi(y) = \sum_{z \in \mathbf{X}} \pi(z)[P^g]_{zy}$.

Remark. This result is well known when there are no transient states (see, e.g., [27, Thm. 1.10, p. 67]).

Proof. Existence. Let R and T denote the recurrent and transient classes under g. By Theorem 1.10 of [27, p. 67], there exists a finite measure $\tilde{\pi}$, unique up to a multiplicative constant, such that

$$\tilde{\pi}(y) = \sum_{z \in R} \tilde{\pi}(z)[P^g]_{zy}.$$

Define the measure π on $\mathbf{X}$ by $\pi(y) = \tilde{\pi}(y)$ for $y \in R$ and $\pi(y) = 0$ otherwise. Then it is easy to check that π solves (3.1), in fact with equality. To prove uniqueness, let π be a solution of (3.1). Iterating (3.1), we obtain for every $n > 0$

$$\pi(y) \geqq \sum_{z \in \mathbf{X}} \pi(z)[(P^g)^n]_{zy} \geqq \sum_{z \in R} \pi(z)[(P^g)^n]_{zy}. \tag{3.2}$$

Again, by Theorem 1.10 of [27, p. 67], there exists $\{\pi(y), y\in R\}$, unique up to a multiplicative constant and independent of n, such that for all $y\in R$, $\pi(y)=\sum_{z\in R}\pi(z)[(P^g)^n]_{zy}$. But this and (3.2) imply that π satisfies $\sum_{z\in T}\pi(z)[(P^g)^n]_{zy}=0$ for all $y\in R$ and all n. Fix some $y\in R$; by (A1), for each $z\in T$ there exists a finite n such that $[(P^g)^n]_{zy}>0$. Thus we conclude $\pi(z)=0$ and the uniqueness is established. □

THEOREM 3.2. *Under* (A1) *the class of stationary policies is weakly complete.*

Proof. First note that $\bar{F}_{x,g}$ is a singleton for any stationary policy g. This follows from the existence of a unique stationary probability, under (A1). Pick any frequency matrix $\zeta\in L_x$ that is achieved, say, by a policy $u\in U$. To establish the theorem we need to show that whenever $\{\bar{f}^t_{x,u}\}_t$ is tight, there exists a policy $g\in U(S)$ so that $\bar{f}_{x,g}=\zeta$. Thus let u be a policy such that $\{\bar{f}^t_{x,u}\}_t$ is tight. Let $\{t_n\}_n$ be an increasing sequence of times (chosen by diagonalization) along which $\lim_{n\to\infty}\bar{f}^{t_n}_{x,u}(y,a):=\bar{f}_{x,u}(y,a)=\zeta(y,a)$ and $\lim_{n\to\infty}\bar{f}^{t_n}_{x,u}(y):=\bar{f}_{x,u}(y)$ exist for all y and $a\in A(y)$. Note that for each y,

$$\bar{f}_{x,u}(y)=\lim_{n\to\infty}\bar{f}^{t_n}_{x,u}(y)=\lim_{n\to\infty}\sum_{a\in A(y)}\bar{f}^{t_n}_{x,u}(y,a) \tag{3.3}$$

and by tightness and the convergence $\bar{f}^{t_n}_{x,u}(y,a)\to\bar{f}_{x,u}(y,a)$, the probability measures $\bar{f}^{t_n}_{x,u}$, $n=1,2,\cdots$ converge weakly (see the Portmanteau theorem in [8]), so that

$$\bar{f}_{x,u}(y)=\lim_{n\to\infty}\sum_{a\in A(y)}\bar{f}^{t_n}_{x,u}(y,a)=\sum_{a\in A(y)}\lim_{n\to\infty}\bar{f}^{t_n}_{x,u}(y,a)=\sum_{a\in A(y)}\bar{f}_{x,u}(y,a). \tag{3.4}$$

Set $\beta^a_y:=\bar{f}_{x,u}(y,a)[\bar{f}_{x,u}(y)]^{-1}$ whenever $\bar{f}_{x,u}(y)\neq 0$. If $\bar{f}_{x,u}(y)=0$ then the β^a_y are chosen arbitrarily but such that $0\leqq\beta^a_y\leqq 1$ for all a, and $\sum_{a\in A(y)}\beta^a_y=1$. By (3.4), $\sum_{a\in A(y)}\beta^a_y=1$ for every y. Define the stationary policy g by $p^g_{a|y}=\beta^a_y$. Then

$$P^g_{xy}=\sum_{a\in A(x)}\beta^a_x P_{xay}. \tag{3.5}$$

Since for every $s>1$ we have $P^u_x\{X_s=y\}=\sum_{z,a}P^u_x\{X_{s-1}=z, A_{s-1}=a\}P_{zay}$, we get after some algebra

$$\bar{f}^t_{x,u}(y)-\frac{1}{t}P^u_x\{X_1=y\}=\sum_{z,a}\bar{f}^t_{x,u}(z,a)P_{zay}-\frac{1}{t}\sum_{z,a}P^u_x\{X_t=z, A_t=a\}P_{zay}. \tag{3.6}$$

Since the left side of (3.6) converges along the sequence t_n to $\bar{f}_{x,u}(y)$, so does the right. Fix y and consider P_{zay} as a σ-finite measure on $\mathbf{X}\times\mathbf{A}$. Applying Fatou's lemma we obtain using (3.6)

$$\bar{f}_{x,u}(y)=\lim_{n\to\infty}\sum_{z,a}\bar{f}^{t_n}_{x,u}(z,a)P_{zay}\geqq\sum_{z,a}\bar{f}_{x,u}(z,a)P_{zay} \tag{3.7}$$

since the last term in (3.6) is bounded by t^{-1}. From (3.5), (3.7) and from the definition of β^a_y we obtain

$$\bar{f}_{x,u}(y)\geqq\sum_z\bar{f}_{x,u}(z)\cdot P^g_{zy}. \tag{3.8}$$

From (3.8) we conclude that $\bar{f}_{x,u}(\cdot)$ is an excessive measure with respect to the transition matrix P^g. It follows from (3.4) that $\{\bar{f}^{t_n}_{x,u}(\cdot)\}_{t_n}$ are tight, and hence $\bar{f}_{x,u}(\cdot)$ is in fact a probability measure. But under (A1), Lemma 3.1 implies that $\bar{f}_{x,u}=\pi^g$. Using the definition of β and g this finally implies that

$$\zeta(y,a)=\bar{f}_{x,u}(y,a)=\bar{f}_{x,u}(y)\cdot\beta^a_y=\pi^g(y)p^g_{a|y}=\bar{f}_{x,g}(y,a) \tag{3.9}$$

by Lemma 1.1. □

From Theorem 3.2 we immediately obtain Corollary 3.3.

COROLLARY 3.3. *Under* (A1) *and* (A2) *the class of stationary policies is complete.*

Combining this with the theorems of § 2 we thus conclude that, under the relevant assumptions, the stationary policies are sufficient for problems OP and COP.

Assumption (A2) is used to guarantee that $\sum_{y,a} \bar{f}_{x,u}(y, a) = 1$ and that $\sum_{a \in A(y)} \beta_y^a = 1$ for every y. Assumption (A0) guarantees the latter; note that it does not imply that $\mathbf{A}$ is finite.

COROLLARY 3.4. *Under* (A1) *and* (A0), *for every policy u and a matrix $\bar{f}_{x,u}(\cdot, \cdot)$, there exists a stationary policy g and a constant $0 \leqq \delta \leqq 1$ such that $\bar{f}_{x,u}(y, a) = \delta \cdot \bar{f}_{x,g}(y, a)$, $y \in \mathbf{X}$, $a \in A(y)$. Under* (A2), $\delta = 1$.

Proof. Following the proof of Theorem 3.2, observe that $\bar{f}_{x,u}(\cdot)$ is an excessive measure due to (3.8) and is thus, by Lemma 3.1, proportional to π^g. But $\bar{f}_{x,u}(\cdot, \cdot)$ is clearly a subprobability measure, i.e., $\sum_{y,a} \bar{f}_{x,u}(y, a) \leqq 1$. Thus by the argument of (3.9), $\bar{f}_{x,u}(y, a) = \delta \cdot \bar{f}_{x,g}(y, a)$, $y \in \mathbf{X}$, $a \in A(y)$. If (A2) holds then it is a probability measure, and $\delta = 1$ by Theorem 3.2. □

Remark. If under every $u \in U(SD)$ there are no transient states then δ in Corollary 3.4 is always strictly positive; moreover, $\bar{f}_{x,u}(y) > \varepsilon(y)$ uniformly in $u \in U$ (see, e.g., [18]).

Before we show that $\bar{f}_{x,\hat{\alpha}}$ depends only on α, we present a simple example that demonstrates the importance of (A2), and shows that a condition such as (A0) is necessary for (A2).

Counterexample 3.5. Countable action space. Consider problem OP with $\mathbf{X} = \{x\}$ and $\mathbf{A} = \{1, 2, \cdots\}$, and let $c(x, a) = 1 + a^{-1}$. Clearly, $X_t = x$ for all t. Under any stationary policy g, $\bar{C}_x(g) > 1$ and $\sum_{a \in A(x)} \bar{f}_{x,g}(x, a) = \bar{f}_{x,g}(x) = 1$. Let u be the nonstationary policy that chooses action $a = t$ at time t. Clearly, we have $\bar{C}_x(u) = 1$, $\bar{f}_{x,u}(x, a) = 0$, $\bar{f}_{x,u}(x) = 1$.

This example demonstrates that even under the unichain assumption, the expected state-action frequencies may not be tight while the expected state frequencies are, and the average expected cost is not necessarily a function of the expected state action frequencies. Moreover, the stationary policies are not complete, and due to the noncompactness of the action space, the cost achieved by some nonstationary policy can be strictly smaller than the cost of any stationary policy. This is in contrast with the case of finite state and action spaces (see Derman [15]).

A counterexample where both (A1) and (A0) hold yet (A2) is not satisfied is presented in Fisher and Ross [17]. They show that indeed without (A2) the stationary policies may be incomplete.

THEOREM 3.6. *Under* (A1) *and* (A2), *$\bar{F}_{x,\hat{\alpha}} = \{\bar{f}_{x,\hat{\alpha}}\}$ is a singleton. Moreover, $\bar{f}_{x,\hat{\alpha}}$ depends only on α and is independent of the realization h.*

Proof. Let $v = \hat{\alpha} = h(\alpha)$ be some *ats* policy with parameter α. Define the stationary policy g by $p^g_{a|y} = \alpha_y^a$. By the strong law of large numbers, g is also an *ats* policy with parameter α. The proof is concluded by showing that $\bar{f}_{x,v}(y, a) = \bar{f}_{x,g}(y, a)$. Since the initial state is fixed, we suppress it in the notation of P and E. Let

$$M_t := \sum_{s=2}^{t} 1\{X_s = y\} - \sum_{s=2}^{t} \sum_{x,a} 1\{X_{s-1} = x, A_{s-1} = a\} P_{xay}.$$

Then for any u, M_t is a P^u martingale and by the stability theorem (e.g., [20, Thm. 2.22])

$$\text{(3.10)} \qquad \lim_{t \to \infty} \frac{1}{t} \left[\sum_{s=2}^{t} 1\{X_s = y\} - \sum_{s=2}^{t} \sum_{x,a} 1\{X_{s-1} = x, A_{s-1} = a\} P_{xay} \right] = 0 \quad P^u \text{ a.s.}$$

Let N be the P^v-null set of ω for which either (3.10) or the convergence in the definition of the *ats* policy v do not hold. Fix $\omega \in \Omega - N$ and an arbitrary increasing sequence of times t_n. Using diagonalization, construct a subsequence s_n to t_n along which for all y and a, $f^{s_n}(y, a)$, $f^{s_n}(y)$ and $f^{s_n}(a|y)$ converge to some limits $f(y, a)$, $f(y)$, and

$f(a\,|\,y)$, respectively. Note that from the definition of the *ats* policy v it follows that $f(y)=\sum_a f(y,a)$ P^v almost surely. For that ω we have from (3.10) for all $y\in\mathbf{X}$:

$$f(y)=\lim_{n\to\infty}\sum_{x,a} f^{(s_n-1)}(x,a)P_{xay}. \tag{3.11}$$

An argument as in (3.7) and (3.8) implies

$$f(y)\geqq\sum_{x,a} f(x,a)P_{xay}. \tag{3.12}$$

From the definition of the *ats* policy v it is easy to see that

$$\alpha_x^a f(x)=f(a\,|\,x)f(x)=f(x,a). \tag{3.13}$$

From (3.12) and (3.13) we obtain, since all terms are nonnegative

$$f(y)\geqq\sum_{x\in\mathbf{X}} f(x)\left[\sum_{a\in A(x)}\alpha_x^a P_{xay}\right]=\sum_{x\in\mathbf{X}} f(x)P_{xy}^g. \tag{3.14}$$

Using the same argument that followed the proof of Corollary 3.4 we obtain for all $y\in\mathbf{X}$:

$$f(y)=\delta(\omega,\{s_n\})\cdot\pi_y^g \tag{3.15}$$

for some constant δ satisfying $0\leqq\delta\leqq 1$. Thus, for all y, z in $\mathbf{X}$,

$$\lim_n\,[f^{s_n}(y)\pi^g(z)-f^{s_n}(z)\pi^g(y)]=0.$$

Since the sequence $\{t_n\}_n$ was arbitrary, we conclude that in fact

$$\lim_t\,[f^t(y)\pi^g(z)-f^t(z)\pi^g(y)]=0$$

and this holds for P^v almost surely. But by the bounded convergence theorem,

$$\lim_t\,[E_v f^t(y)\pi^g(z)-E_v f^t(z)\pi^g(y)]=\lim_t\,[\bar f^t_{x,v}(y)\pi^g(z)-\bar f^t_{x,v}(z)\pi^g(y)]=0.$$

By assumption $\{\bar f^t_{x,v}\}_t$ is tight. Fix any subsequence $\{r_n\}_n$ such that $\bar f^{r_n}_{x,v}\to\bar f_{x,v}$. Then $\bar f_{x,v}(y)\pi^g(z)=\bar f_{x,v}(z)\pi^g(y)$. However, the only probability measure that solves this equation is $\bar f_{x,v}=\pi^g$, and we conclude that $\bar f^t_{x,v}\to\pi^g$. From the definition of the *ats* policy v and the bounded convergence theorem, we have

$$\overline{\lim_{t\to\infty}}\,E_v|\alpha_y^a-f^t(y,a)[f^t(y)]^{-1}|=0.$$

Thus, using the bounded convergence theorem and the tightness (A2),

$$\lim_{t\to\infty}\bar f^t_{x,v}(y,a)=\lim_{t\to\infty}E_v f^t(y)\cdot\frac{f^t(y,a)}{f^t(y)}=\lim_{t\to\infty}\bar f^t_{x,v}(y)\cdot\alpha_y^a=\alpha_y^a\pi^g(y) \tag{3.16}$$

for all a, y. Finally, Lemma 1.1 implies $\bar f_{x,g}(y,a)=\alpha_y^a\cdot\pi^g(y)=\bar f_{x,v}(y,a)$. Since π^g depends only on α and not on $\hat\alpha$ this concludes the proof. □

Combining Theorem 3.2 and Theorem 3.6 it follows that the completeness is determined by the α only, so that complete classes can be easily generated.

THEOREM 3.7. *Under* (A1) *and* (A2), *for any realization* $h:S_\alpha\to U$, $U^h(ats)$ *is complete.*

4. Tightness. The issue of tightness for Markov decision processes has been investigated extensively. It is easy to see that, in general, unless the sets $A(x)$ are finite (compact), (A2) need not hold. In the compact case, Lemma 4.1 provides a simple sufficient condition for (A2). We describe briefly several approaches that provide sufficient conditions for tightness in this compact case (i.e., under (A0)).

If compactness of the actions is not assumed, we can usually construct a policy u for which (A2(u)) does not hold, so that (A2) will not hold. However, since the tightness appears in connection with the optimization problems, we derive conditions on the cost functions that guarantee that the search for optimal policy can be restricted to policies satisfying (A2(u)). This extends the results of §§ 2.2 and 2.3 to cases where (A2) does not hold.

LEMMA 4.1. *Under* (A0) *and* (A1), (A2) *implies* (A2*). *If in addition there are no transient states, then* A2 *is equivalent to* (A2*).

Proof. It is shown in the proof of Lemma 7.3 of [10] that, under (A0) and (A1) and when there are no transient states, (A2*) implies that for each state x and policy u the expected frequencies $\{\bar f^t_{x,u}(y)\}_t$ are tight. To see that the converse holds, assume (A0) and (A1) and let g_i be a sequence of policies in $U(SD)$ such that the sequence of corresponding invariant distributions π_i is not tight. Clearly, $\bar f^t_{x,g_i}(y)\rightarrow\pi_i(y)$ for all x, y. Thus we can find an increasing sequence $\{t_i\}$ and construct a policy u where $u_t(\cdot\,|\,H_{t-1},X_t)=g_i(x_t)$ for $t_i<t\leqq t_{i+1}$ such that $\bar f^{t_i}_{x,u}\rightarrow\bar f_{x,u}$, and $\sum_y\bar f_{x,u}(y)<1$. Thus it suffices to show that $\{\bar f^t_{x,u}(y)\}_t$ is tight if and only if $\{\bar f^t_{x,u}(y,a)\}_t$ is tight.

By definition, $\{\bar f^t_{x,u}(y)\}_t$ is tight if and only if for any $\varepsilon>0$ there exists a compact (finite) set $K(\varepsilon)\subset\mathbf{X}$ such that $\sum_{y\in K(\varepsilon)}\bar f^t_{x,u}(y)>1-\varepsilon$, and similarly for $\{\bar f^t_{x,u}(y,a)\}_t$. Given $K(\varepsilon)$, let $K'(\varepsilon):=\{(y,a)\colon y\in K(\varepsilon),\,a\in A(y)\}$. Then $K'(\varepsilon)$ is compact and since

$$\bar f^t_{x,u}(y)=\sum_{a\in A(y)}\bar f^t_{x,u}(y,a) \tag{4.1}$$

we have $\sum_{(y,a)\in K'(\varepsilon)}\bar f^t_{x,u}(y,a)=\sum_{(y)\in K(\varepsilon)}\bar f^t_{x,u}(y)>1-\varepsilon$. To prove the converse, given $K'(\varepsilon)\subset\mathbf{X}\times\mathbf{A}$ let $K(\varepsilon):=\{y\colon(y,a)\in K'(\varepsilon)$ for *some* $a\in A(y)\}$. The same argument now concludes the proof. □

Assumption (A2*) is quite common in the literature on controlled Markov chains with a countable state space, and sufficient conditions are available. Borkar [10, § III] shows that (A2*) is equivalent to the time between visits to some recurrent state being uniformly integrable under all $u\in U(SD)$. The whole § IX in [10] is then devoted to different sufficient conditions for that uniform integrability. Hordijk [21] presents several sufficient conditions for (A2*), in terms of the measures $P^g_{x,K}:=\sum_{y\in K}P^g_{xy}$;

(i) The set of probability measures $\{P(X_2=\cdot\,|\,X_1=x)\colon x\in\mathbf{X},\,g\in U(S)\}$ is tight [21, Lemma 10.3, § 10].

(ii) Given any $\varepsilon>0$ there exist a finite set $K(\varepsilon)$ and an integer $N(\varepsilon)$ such that for all $x\in\mathbf{X}$ and $g\in U(S)$,

$$[(P^g)^{N(\varepsilon)}]_{x,K(\varepsilon)}\geqq 1-\varepsilon.$$

(iii) The simultaneous Doeblin condition. There exist a finite set K, a positive integer n, and a positive real number c such that $[P^g]^n_{x,K}\geqq c$ for all $x\in\mathbf{X}$ and all $g\in U(S)$ [21, § 11.1].

Two other assumptions that are equivalent to (iii) above (and are thus sufficient for (A2*)) are presented in Theorem 11.3 of [21]. To formulate these conditions, denote

$${}_AP^{g,t}_{x,B}:=P^g\{X_t\in B,\,X_s\notin A,\,1<s<t\,|\,X_1=x\},\qquad m_g(x,A):=\sum_{t=2}^{\infty}{}_AP^{g,t}_{x,\mathbf{X}}.$$

(iv) There exist a finite set K, $c>0$, and n such that $\sum_{s=2}^{n}{}_KP^{g,s}_{x,K}\geqq c$ for all $x\in\mathbf{X}$ and $g\in U(S)$.

(v) There exist a finite set K and a real number b such that for all $x\in\mathbf{X}$ and $g\in U(S)$, $m_g(x,K)\leqq b$.

In the absence of tightness, it may be possible to restrict the optimization problem to a subclass of policies under which tightness holds, if the structure of the costs makes it unprofitable to use nontight policies (see also Borkar [10]).

LEMMA 4.2. *Assume there exists a sequence of increasing compact (i.e., finite) subsets* K_i *of* $\mathbf{X}\times\mathbf{A}$ *such that* $\bigcup_i K_i=\mathbf{X}\times\mathbf{A}$, *and such that the cost function* $c(y, a)$ *satisfies*

$$\liminf_{i\to\infty}\{c(y, a);\ (y, a)\notin K_i\}=\infty. \tag{4.2}$$

Then for any policy u *such that* $\bar{C}_x(u)<\infty$ (*or* $\underline{C}_x(u)<\infty$), *the frequencies* $\{\bar{f}^{t}_{x,u}(\cdot,\cdot)\}_t$ *are tight.*

Proof. By (4.2), $c(x, a)$ is bounded below, say by B. Assume $\{\bar{f}^{t}_{x,u}(\cdot,\cdot)\}$ is not tight. Then there exists some $\varepsilon>0$ and an increasing sequence $\{t_l\}$ such that $\sum_{(y,a)\notin K_l}\bar{f}^{t_l}_{x,u}(y, a)>\varepsilon$. Let $c_j:=\inf\{c(y, a);\ (y, a)\notin K_j\}$. Clearly, $\bar{C}^{t_l}_x(u)\geqq c_j\varepsilon-|B|$. But by (4.2) $\lim_{j\to\infty}c_j=\infty$, and hence $\bar{C}_x(u)=\infty$, contradicting the hypotheses. The proof using $\underline{C}$ is identical. □

A complete class of policies (or even a weakly complete class of policies) may thus be sufficient even when the tightness assumption (A2) does not hold.

THEOREM 4.3. *Assume* (A1) *and consider the problem of minimizing* $\bar{C}_x(u)$. *Let* U' *be a weakly complete class such that* (2.4) *holds for each* $v\in U'$. *If* $c(\cdot,\cdot)$ *satisfies the conditions of Lemma* 4.2, *then* U' *is sufficient for* OP.

Note that the stationary policies are in fact weakly complete (Corollary 3.4) and, under (A3(g)), satisfy (2.4).

Proof. From Lemma 4.2 we conclude that if $\{\bar{f}^{t}_{x,u}\}_t$ is not tight (so that for some limit point $\sum_{y,a}\bar{f}_{x,u}(y, a)<1$), then necessarily $C_x(u)=\infty$. Thus we may limit the optimization to policies u for which $\{\bar{f}^{t}_{x,u}\}_t$ is tight, so that $\sum_{y,a}\bar{f}_{x,u}(y, a)=1$. By (4.2) $c(\cdot,\cdot)$ is bounded from below, and hence by Corollary 2.5 U' is sufficient. □

Similarly for the constrained problem COP, we may relax (A2) in Theorem 2.8.

THEOREM 4.4. *Assume* (A1) *and consider problem* COP. *Let* U' *be a weakly complete class such that* (2.4) *holds for each* $v\in U'$. *If either* $c(\cdot,\cdot)$ *or* $d^k(\cdot,\cdot)$, *some* k *satisfies the conditions of Lemma* 4.2, *then* U' *is sufficient for* COP.

Proof. The proof is similar to that of Theorem 4.3. □

Next we present another method that provides conditions for sufficiency in cases that the tightness does not hold. It is a generalization of conditions that Borkar [9] introduced for the case of instantaneous cost that depends only on the state. Following [9], $c(\cdot,\cdot)$ is said to be "V-almost monotone" if there exists a collection of compact sets $\{K_i\}_i$ as in Lemma 4.2 such that $\liminf_{i\to\infty}\{c(y, a);\ (y, a)\notin K_i\}\geqq V$.

LEMMA 4.5. *Assume* (A0) *and* (A1) *and let* U' *be a weakly complete class of policies such that every* $u\in U'$ *satisfies* (2.4). *If* $c(\cdot,\cdot)$ *is V-almost monotone and* $C_x(u')\leqq V$, *some* $u'\in U'$, *then* U' *is sufficient for* OP.

Proof. Note that $c(\cdot,\cdot)$ is bounded below. Consider first the minimization of $\bar{C}_x$, fix an arbitrary v, and note that if $C_x(v)\geqq V$ then we are done. Thus assume $C_x(v)<V$ and let t_n be a subsequence such that $C_x(v)=\lim_{n\to\infty}\sum_{y,a}\bar{f}^{t_n}_{x,v}(y, a)c(y, a)$ and $\bar{f}^{t_n}_{x,v}$ converges to some $\bar{f}_{x,v}$. By Corollary 3.4, there exists a $g\in U(S)$ such that $\delta\bar{f}_{x,g}=\bar{f}_{x,v}$ for some $0\leqq\delta\leqq 1$. By completeness there exists a $u\in U'$ such that $\delta\bar{f}_{x,u}=\bar{f}_{x,v}$. Let ε_i be such that $\inf\{c(y, a);\ (y, a)\notin K_i\}\geqq V-\varepsilon_i$. For every i we have

$$\begin{aligned} C_x(v) &= \lim_{n\to\infty}\left[\sum_{(y,a)\in K_i}\bar{f}^{t_n}_{x,v}(y, a)c(y, a)+\sum_{(y,a)\notin K_i}\bar{f}^{t_n}_{x,v}(y, a)c(y, a)\right] \\ &\geqq \sum_{(y,a)\in K_i}\bar{f}_{x,v}(y, a)c(y, a)+\underline{\lim}_{n\to\infty}\sum_{(y,a)\notin K_i}\bar{f}^{t_n}_{x,v}(y, a)c(y, a), \end{aligned}$$

$$\underline{\lim}_{n\to\infty} \sum_{(y,a)\notin K_i} \bar{f}^{t_n}_{x,v}(y,a)c(y,a) \geqq (V-\varepsilon_i) \underline{\lim}_{n\to\infty} \sum_{(y,a)\notin K_i} \bar{f}^{t_n}_{x,v}(y,a)$$

$$= (V-\varepsilon_i) \underline{\lim}_{n\to\infty} \left[1 - \sum_{(y,a)\in K_i} \bar{f}^{t_n}_{x,v}(y,a)\right].$$

Using the fact that each term on the right converges to $\bar{f}_{x,v}(y,a)$ and that $\bar{f}_{x,v} = \delta \bar{f}_{x,u}$, we get

$$C_x(v) \geqq \delta \sum_{(y,a)\in K_i} \bar{f}_{x,u}(y,a)c(y,a) + (V-\varepsilon_i)\left[1 - \delta \sum_{(y,a)\in K_i} \bar{f}_{x,u}(y,a)\right].$$

By taking $i\to\infty$ we obtain $C_x(v) \geqq \delta C_x(u) + V(1-\delta)$ or $\delta C_x(v) \geqq \delta C_x(u) + (1-\delta)\times (V - C_x(v))$. Since the last term is positive, $C_x(v) \geqq C_x(u)$ that establishes the proof. The proof for $\underline{C}$ is identical. □

In the following lemma we apply the method of Lemma 4.5 in order to generalize Theorem 4.4.

LEMMA 4.6. *Assume* (A0) *and* (A1). *Let U' be a weakly complete class of policies such that every $u \in U'$ satisfies* (2.4) *for c and d^k, $1 \leqq k \leqq K$. Assume $c(\cdot,\cdot)$ is V_0-monotone and $d^k(\cdot,\cdot)$ is V_k-monotone, $1\leqq k \leqq K$. If there exists a policy $u' \in U'$ such that $C_x(u') \leqq V_0$ and $D^k_x(u') \leqq V_k$, $1 \leqq k \leqq K$, then U' is sufficient for* COP.

Proof. The proof is the same for both $\bar{C}$ and $\underline{C}$. Note that without loss of generality, we may take the sets K_i to be the same for all costs c and d^k. Fix $v \in U'$ and note that if $C_x(v) \geqq V_0$ or for some k $D^k_x(v) > V_k$ then we are done. Thus assume $D^k_x(v) \leqq V_k$ for $k = 1, \cdots, K$ and $C_x(v) \leqq V_0$. Choose a subsequence $\{t_n\}_n$ such that $C_x(v) = \lim_{n\to\infty} \sum_{y,a} \bar{f}^{t_n}_{x,v}(y,a)c(y,a)$ and such that $\lim_{n\to\infty} \sum_{y,a} \bar{f}^{t_n}_{x,v}(y,a)d^k(y,a)$, $1 \leqq k \leqq K$, and $\lim_{n\to\infty} \bar{f}^{t_n}_{x,v}$ exist. Then for $1 \leqq k \leqq K$, $D^k_x(v) \geqq \overline{\lim}_{n\to\infty} \sum_{y,a} \bar{f}^{t_n}_{x,v}(y,a)d^k(y,a)$ and by choosing u as in Lemma 4.5 we obtain by the same argument $\delta C_x(v) \geqq \delta C_x(u) + (1-\delta)(V_0 - C_x(v))$ and $\delta D^k_x(v) \geqq \delta D^k_x(u) + (1-\delta)(V_k - D^k_x(v))$. Hence $C_x(v) \geqq C_x(u)$ and $V_k \geqq D^k_x(v) \geqq D^k_x(u)$ that concludes the proof. □

5. Achievable frequencies. In this section we describe the geometry of L_x. For the case of finite state and action spaces Derman [15] has shown that under (A1), L_x is closed and convex, and its extreme points correspond to policies in $U(SD)$. In Theorem 5.1 we extend this result to the countable space. Let co B denote the convex hull of the set B, and $\overline{\text{co}}$ B, its closed convex hull. Let η be a function from the integers onto all pairs (x,a) and fix $0 < \mu < 1$. Define a metric d on the set of subprobability measures on $\mathbf{X}\times\mathbf{A}$.

$$d(\zeta_1,\zeta_2) := \sum_{j=1}^{\infty} |\zeta_1(\eta[j]) - \zeta_2(\eta[j])|\mu^j. \tag{5.1}$$

We will use henceforth the product topology induced by this metric. Throughout this section we assume that

(A1′) The state space forms a single *positive* recurrent class under any policy $g \in U(S)$.

To prove Theorem 5.1, we need to introduce PTS ("policy time sharing") policies [2]. A PTS policy is specified through the stationary policies u_i, $i = 1, 2, \cdots, l$, a state z, and by an l-dimensional vector parameter $\alpha = \{\alpha_1, \alpha_2, \cdots, \alpha_l\}$, where $\alpha_i \geqq 0$ and $\sum_i \alpha_i = 1$. Fix a state z that, due to (A1′) is positive recurrent under each u_i. Call the period between two consecutive visits to states z a cycle. A PTS policy v with parameter α is any policy that uses a fixed u_i during each cycle, and for which the relative number of cycles during which u_i is used converges to α_i, P^v almost surely, $i=1, 2, \cdots, l$. Such a policy is denoted $\hat{\alpha}$. It follows immediately from the results of [1] and [3] that for any initial state x, $\bar{F}_{x,\hat{\alpha}}$ is a singleton, and

$$\bar{f}_{x,\hat{\alpha}} = \sum_{i=1}^{l} \gamma_i(\alpha)\bar{f}_{x,u_i}, \tag{5.2}$$

where $\gamma_i = \alpha_i\tau_i[\sum_{j=1}^{l} \alpha_j\tau_j]^{-1}$ and τ_i is the mean recurrence time of state z under u_i.

THEOREM 5.1. *Under* (A1′) *and* (A2), $L_x = L_x(S)$ *is compact. Moreover,* $L_x = \mathrm{co}\,\{L_x(D)\} = \overline{\mathrm{co}}\,\{L_x(D)\}$.

Proof. By Corollary 3.3, $L_x(S) = L_x$. To prove compactness, let $\{\xi_n\}_n \subset L_x$. Using diagonalization, choose a subsequence $\{\xi_{n_i}\}_i$ that (for notational convenience) is denoted $\{\zeta_n\}_n$, such that $\zeta_n(x, a) \to \zeta(x, a)$ for some ζ, for all x, a. ζ_n may all be considered measures over $\mathbf{X} \times \mathbf{A}$, and $\zeta_n(x, a) \to \zeta(x, a)$, where ζ is (possibly subprobability) measure. Our aim is to find a policy u that achieves ζ. By (A2) this implies that ζ is a probability measure.

By Corollary 3.3 there exists a stationary policy g_i that achieves ζ_i. Let $\varepsilon_i := d(\zeta, \zeta_i)$, so that $\lim_{n\to\infty} \varepsilon_n = 0$. Consider the nonstationary policy u, that uses g_1 until the time $t_1 := \min\{t\colon d(\zeta_1, \bar{f}^{\,t}_{u,x}) \leqq \varepsilon_1\}$, and uses g_i until between t_{i-1} and $t_i := \min\{t > t_{i-1}\colon d(\zeta_i, \bar{f}^{\,t}_{u,x}) \leqq \varepsilon_i\}$. The fact that $t_n < \infty$ can be proved by induction using the following fact. Suppose the policy u uses g_n from time s onward, and let $\chi_s(z) = P^u(X_s = z \mid X_1 = x)$. Then

$$\bar{f}^{\,t}_{x,u}(y, a) = \frac{s}{t}\bar{f}^{\,s}_{x,u}(y, a) + p^{g_n}_{a|y}\frac{t-s}{t}\sum_{z\in\mathbf{X}} \chi_s(z)\left(\sum_{r=1}^{t-s} [P^{g_n}]^r_{zy}\right),$$

where P^{g_n} is the transition matrix under g_n. It then follows easily that $\lim_{t\to\infty} \bar{f}^{\,t}_{x,u}(y, a) = \zeta_n$.

Thus $d(\zeta, \bar{f}^{\,t_n}_{x,u}) \leqq d(\zeta, \zeta_n) + d(\zeta_n, \bar{f}^{\,t_n}_{x,u}) \leqq 2\varepsilon_n$ and we obtain along the subsequence $\{t_n\}_n$, $\bar{f}_{x,u} = \zeta$. By (A2) ζ is a probability measure, so that L_x is closed and sequentially compact, hence compact.

To prove the convexity, recall (the first part of the proof) that $L_x = L_x(S)$. Suppose $\zeta = \beta\bar{f}_{x,u_1} + (1-\beta)\bar{f}_{x,u_2}$ for u_1, $u_2 \in U(S)$. A PTS policy u such that $\bar{f}_{x,u} = \zeta$ is obtained by setting $\alpha_1 := (\beta/\tau_1)(\beta/\tau_1 + (1-\beta)/\tau_2)^{-1}$, and $\alpha_2 = 1 - \alpha_1$ (this follows from (5.2)).

Since L_x is compact and convex in $\mathbb{R}^\infty$, by the Krein–Milman theorem it is the convex hull of its extreme points. Next we show that all extreme points of L_x correspond to deterministic stationary policies. Let g bé a stationary nondeterministic policy. Then for some state $z \in \mathbf{X}$ and actions a_1 and a_2 in $A(z)$, the probability α_i to use action a_i under the policy g is strictly positive. Consider the stationary policies u_i that coincide with g in all states except for state z. In state z policy u_i uses action a_i with probability $\alpha_1 + \alpha_2$. Then according to (5.2), the PTS policy $\hat{\alpha}$ that switches in state z between u_1 and u_2 achieves $\bar{f}_{x,g} = \gamma\bar{f}_{x,u_1} + (1-\gamma)\bar{f}_{x,u_2}$. Therefore $\bar{f}_{x,g}$ is not an extreme point in L_x, and since for every policy u there is a $g \in U(S)$ with $\bar{f}_{x,u} = \bar{f}_{x,g}$ this concludes the proof. □

Theorem 5.1 enables us to strengthen theorem 2.6 as follows.

COROLLARY 5.2. *Assume* (A1′) *and* (A2) *under the uniform integrability assumption, or under the assumption that c is bounded from below, the class of deterministic policies is sufficient for problem* OP (*with C defined through either* (1.2a) *or* (1.2b)).

Proof. By Lemma 2.3, the cost of a stationary policies has the representation (2.4). An argument as in the proof of Theorem 5.1 then shows that the cost of any nondeterministic policy is a convex combination of the costs of two other stationary policies. □

Another conclusion from Theorem 5.1 is that under (A1′) and (A2) the state frequencies are bounded from below by a positive (state dependent) constant, uniformly in the policy.

COROLLARY 5.3. *Under* (A1′) *and* (A2), *for each* $y \in \mathbf{X}$ *there exists a constant* $\Delta(y) > 0$ *such that* $\bar{f}_{x,u}(y) > \Delta(y)$ *for all policies* $u \in U$.

Proof. Suppose the claim does not hold. Then there exists a sequence g_n of stationary policies and some state z such that $\lim_{n\to\infty} \bar{f}_{x,g_n}(z) = 0$. We can then construct a subsequence n_k along which $\lim_{n\to\infty} \bar{f}_{x,g_n}(y)$ exists for all $y \in \mathbf{X}$. Using Theorem 5.1 and (3.3), (3.4) there exists some stationary policy g achieving this limit, hence $\bar{f}_{x,g}(z) = 0$, which contradicts (A1). □

Remark. Fisher [18] showed that if the state space forms a single *positive* recurrent class when using any deterministic policy $g \in U(SD)$ then (A1′) holds. He then obtained the same result as in Corollary 5.3 using only the weaker condition (A0) instead of (A2).

Finally, we use Theorem 5.1 to strengthen Theorem 2.8. Theorem 2.8 states that we may restrict our search for optimal policy for COP to the stationary policies. But it does not say that an optimal policy exists. We show that this is indeed the case.

COROLLARY 5.4. *Consider problem* COP. *Assume* (A1′) *and* (A2) *and either*

(i) *Both* $c(y, a)$ *and* $d^k(y, a)$ *are bounded from below*; *or*

(ii) *Both* $c(y, a)$ *and* $d^k(y, a)$ *are uniformly integrable with respect to* $\{\bar{f}_{x,g}\}$, $g \in U(S)$.

If there is any feasible policy then there exists an optimal stationary policy.

Proof. According to Theorem 2.8 we may restrict to the stationary policies in searching for optimality. We first show that $C_x(\cdot)$ and $D_x^k(\cdot)$ are lower semicontinuous functions of the frequencies $\bar{f}_{x,g}$. Let $\{\zeta_n\}$ be a sequence of frequencies, achieved, say, by the stationary policies $\{g_n\}$ (i.e., $\zeta_n(\cdot,\cdot) = \bar{f}_{x,g_n}(\cdot,\cdot)$) converging to ζ. According to Theorem 5.1 there exists a stationary policy g such that $\zeta = \bar{f}_{x,g}$. Under (i) this implies by Fatou's lemma (using $c(\cdot,\cdot)$ as a measure) that the cost function $C_x(g)$ satisfies

$$\begin{aligned} C_x(g) &= \sum_{y,a} \zeta(y, a)c(x, a) = \sum_{y,a} \lim_{n\to\infty} \zeta_n(y, a)c(y, a) \\ &\leqq \lim_{n\to\infty} \sum_{y,a} \zeta_n(y, a)c(y, a) = \lim_{n\to\infty} C_x(g_n) \end{aligned} \tag{5.3}$$

and similarly

$$D_x^k(g) = \sum_{y,a} \lim_{n\to\infty} \zeta_n(y, a)d^k(y, a) \leqq \lim_{n\to\infty} D_x^k(g_n), \tag{5.4}$$

which establishes the lower semicontinuity for the case (i). If (ii) is assumed, then in fact we have continuity. To see that, note that the compactness of $L_x(S)$ (Theorem 5.1) implies by Prohorov's theorem that $\{\zeta_n\}$ are tight, hence converge weakly. As in the proof of Lemma 2.2, consider now $c(y, a)$ and $d^k(y, a)$ as "random variables" on the space $\mathbf{X} \times \mathbf{A}$. The weak convergence of $\{\zeta_n\}$ implies the convergence of $c(\cdot,\cdot)$ and $d^k(\cdot,\cdot)$ in distribution, and combining it with (ii) we obtain $C_x(g) = \lim_{n\to\infty} C_x(g_n)$ and $D_x^k(g) = \lim_{n\to\infty} C_x^k(g_n)$.

We thus have lower semicontinuity under either (i) or (ii). This implies that the set $\Pi_V := \{\mu\colon \mu \in L(S), D_x^k(\mu) \leqq V^k, 1 \leqq k \leqq K\}$ is compact, since it is obtained as the intersection of the compact set $L(S)$ and the inverse map of the closed sets $(-\infty, V_k]$. Finally, by the lower semicontinuity of $C_x(\cdot)$ on Π_V we conclude that $C_x(\cdot)$ achieves its minimum on Π_V, i.e. there exists an optimal stationary policy for COP. □

6. Application to a queueing system. In this section we apply Theorems 2.6 and 3.2 to investigate a constrained problem in the following discrete-time queueing model. At time t, M_t^k customers arrive to queue k, $1 \leqq k \leqq K$. Each input stream is received in an infinite capacity buffer. Arrival vectors $M_t = \{M_t^1, \cdots, M_t^K\}$ are independent from slot to slot, form a renewal sequence with finite means λ_k. During a time slot $(t, t+1)$ a customer from any class k, $1 \leqq k \leqq K$ may be served, according to some policy, which is a prespecified dynamic priority assignment. If served, with probability μ_k it completes its service and leaves the system; otherwise it remains in its queue. A generic element of the state is given by $x = \{x^1, x^2, \cdots, x^K\}$ and it represents a K-dimensional vector of the different queue sizes. Altman and Shwartz [1], [3] solve a problem with constraints on the average sizes of several queues. They find an optimal nonstationary time sharing policy, using a linear program. The recurrence properties of this system as well as bounds and representations for average cost functionals for general cost functions are obtained in Makowski and Shwartz [25].

Below we present conditions for the completeness of stationary policies , and the existence of optimal stationary policies for COP with several constraints. Sufficiency is proved for costs that are nonlinear in the queue sizes. We then solve the general constrained problem with linear costs (generalizing [1], [3], [26]). Throughout we restrict to nonidling policies; using coupling (as in [11]) it can be shown that when the costs are positive and increasing (in the number of customers), idling leads to no improvement.

6.1. Completeness and sufficiency of stationary policies. We first show that (A1) and (A2) hold. We assume the standard stability condition on the traffic intensity $\rho := \sum_{k=1}^K \lambda_k/\mu_k < 1$. This is a sufficient condition for (A1) (see [25] or [26]). In order to show that (A2) is satisfied, we use Lemma 4.2. Let $c(x, a) = \sum_{k=1}^K x^k/\mu_k$. The average cost is then finite and does not depend on the policy (this follows from the μc rule [5], [6], [11]). Therefore (with the obvious choice of K_i) all conditions of Lemma 4.2 are satisfied, and (A2) holds. Hence we obtain, using Theorem 2.8, Corollary 3.3, and Corollary 6.1.

COROLLARY 6.1. *Under the foregoing assumptions on the queueing model, the stationary policies are complete. If $c(x, a)$ and $d^k(x, a)$ are bounded below then the stationary policies are sufficient for* COP.

If $c(x, a)$ and $d^k(x, a)$ are not bounded from below then the stationary policies are still sufficient for COP, provided $\{c(X_s, A_s)\}_s$ and $\{d^k(X_s, A_s)\}_s$ are uniformly integrable with respect to P^u for each policy u (Theorem 2.8). In [25], Makowski and Shwartz give the following sufficient conditions (P1) and (P2) for the uniform integrability. For any K-dimensional vector x let $|x|$ denote $\sum_{k=1}^K |x^k|$:

(P1) There exists an integer $\gamma > 1$ such that $E[|M_t|^\gamma] < \infty$ and $E[|X_1|^\gamma] < \infty$.

Note that both expectations are independent of the policy.

(P2) There exist $0 < \delta < \gamma - 1$ and L such that $|c(x, a)| \leqq L(1 + |x|^\delta)$.

These results establish that the search for optimal, or ε-optimal policies may be restricted to the stationary policies. This allows the application of steady-state analysis, of the type used in queueing theory, to problems OP and COP.

6.2. Solving COP with linear cost functions. Consider the linear cost function $c(x, a) := \sum_{k=1}^K c_k x^k$ and $d^i(x, a) = \sum_{k=1}^K d_k^i x^k$ for $1 \leqq i \leqq M$, where c_k and d_k^i are nonnegative constants. This COP problem was solved in [1] and [3] for the case $M = 1$, and for the case $d_k^i = \delta_i(k)$ and $M < K$ using "PTS" policies over the set of $l = K!$ priority policies g_i. It is shown [1] and [3] that under the condition $\rho < 1$ and $E|X_1| < \infty$, all the τ_i (defined below (5.2)) are equal, and the cost under $\hat{\alpha}$ is given by

$$C_x(\hat{\alpha}) = \sum_{i=1}^{l} \alpha_i C(g_i) \qquad (6.1)$$

with a similar linear expression for $D_x^k(\hat{\alpha})$. Denote by $\hat{\beta}$ the optimal policy among all PTS policies for problem COP. From (6.1) it follows that $\hat{\beta}$ can be obtained by solving the following linear program:

$$(6.2)\quad \text{(LP)}\qquad \text{Find } \alpha \text{ that minimizes } \sum_{i=l}^{l} \alpha_i C(g_i)$$

$$\text{subject to } \sum_{i=1}^{l} \alpha_i D^k(g_i) \leqq V_k, \quad 1 \leqq k \leqq K, \quad \sum_{i=1}^{l} \alpha_i = 1, \quad \alpha_i \geqq 0$$

$$\text{for } 1 \leqq i \leqq l.$$

Based on Theorem 2.8(ii) we show in the following theorem that $\hat{\beta}$ is in fact overall optimal.

THEOREM 6.2. *The* PTS *policy obtained by solving* LP *is optimal for* COP.

Proof. Following [3] we define the average size of queue k by

$$(6.3)\qquad \bar{X}_x^k(u) := \overline{\lim_{t\to\infty}} \frac{1}{t} E_u\left[\sum_{s=1}^{t} X_t^k \,|\, X_1 = x\right].$$

Consider the class U' of all policies satisfying

$$(6.4)\qquad C_x(u) = \sum_{k=1}^{K} c_k \bar{X}_x^k(u) \quad \text{and} \quad D_x^j(u) = \sum_{k=1}^{K} d_k^j \bar{X}_x^k(u), \quad 1 \leqq j \leqq M.$$

Note that $U(S) \subset U'$ and $U(PTS) \subset U'$ (this is obtained by applying Lemma 2.3 to compute $\bar{X}_x^k(u)$). According to Theorem 2.8(ii) $U(S)$ is sufficient hence U' must be sufficient. Reference [3] shows that PTS policies are "Pareto optimal" in the following sense. For any policy u there exists a PTS policy w such that $\bar{X}_x^k(w) \leqq \bar{X}_x^k(u)$, $1 \leqq k \leqq K$. This implies that $\hat{\beta}$ is optimal over U', and since U' is sufficient, this implies that $\hat{\beta}$ is overall optimal. □

This result illustrates the usefulness of the present approach. There are several results reducing optimization problems for queues to computable problems (such as linear programs). However, the optimization is usually carried out over a class of policies that is smaller than U' above (e.g., in [19] the optimization is carried out over the class of "steady state" policies). Results on sufficiency then allow to conclude optimality over the class U of *all* policies.

7. Second application: a linear program formulation for COP. Below we present a linear program that we show to be equivalent to COP. Such linear programs have been introduced for the case of finite state and action spaces (e.g., Derman [15] and Hordijk and Kallenberg [22]). In the finite case these are the most important method to compute optimal policies for COP (an alternative linear program is described in [2]). We use a different approach by which we obtain a similar linear program for the countable case. Naturally, we cannot expect to find explicit solutions for COP using an infinite-dimensional linear program, but this approach can be used to shed some light on the structure of optimal solutions for COP. Consider the LP.

Find $\{z^*(y, a)\}_{y,a}$ that minimizes $C(z) := \sum_{y,a} c(y, a) z(y, a)$ subject to

$$(7.1a)\qquad \sum_{y,a} z(y, a) P_{yav} \leqq \sum_a z(v, a), \qquad v \in \mathbf{X},$$

$$(7.1b)\qquad \sum_{y,a} d^k(y, a) z(y, a) \leqq V_k, \qquad 1 \leqq k \leqq K,$$

$$(7.1c)\qquad \sum_{y,a} z(y, a) = 1, \qquad z(y, a) \geqq 0.$$

THEOREM 7.1. *Assume* (A1) *and* (A2) *and assume either that* (i) *c and* d^k *are bounded below, and* (A3(*g*)) *holds for all stationary g, with respect to both c and* d^k, $1 \leqq k \leqq K$, *or that* (ii) $\{c(X_s, A_s)\}_s$ *and* $\{d^k(X_s, A_s)\}_s$, $1 \leqq k \leqq K$ *are uniformly integrable under* P^u *for all* $u \in U$.

(i) *If the stationary policy w is feasible for* COP, *then* $\{z(y, a)\}$ *satisfies* (7.1), *where*

$$z(y, a) = \pi_y^w \cdot p_{a|y}^w. \tag{7.2}$$

(ii) *If g is an optimal stationary policy for* COP *then there exists an optimal solution for* LP *satisfying*

$$z^*(y, a) = \pi_y^g \cdot p_{a|y}^g. \tag{7.3}$$

(iii) *Conversely, let* $\{z(y, a)\}$ *satisfy* (7.1). *Then the policy w is feasible for* COP, *where*

$$p_{a|y}^w = \frac{z(y, a)}{\sum_{a \in A(y)} z(y, a)} \tag{7.4}$$

whenever the denominator is nonzero, and otherwise $p_{a|y}^w$ *are chosen arbitrarily but such that* $p_{\cdot|y}^w$ *is a probability measure.*

(iv) *If* z^* *solves* LP, *then the stationary policy g is optimal for* COP, *where*

$$p_{a|y}^g = \frac{z^*(y, a)}{\sum_{a \in A(y)} z^*(y, a)} \tag{7.5}$$

whenever the denominator is nonzero, and otherwise $p_{a|y}^g$ *are chosen arbitrarily but such that* $p_{\cdot|y}^g$ *is a probability measure.*

Proof. To prove (i) we note that $z(y, a)$ as defined in (7.2) satisfies (7.1c) since π^w and $p_{\cdot|y}^w$ are probability measures. Next we note that $z(a, y) = \bar{f}_{x,w}(a, y)$, thus (7.1b) is satisfied since its left side is equal to $D^k(g)$ by Lemma 2.3. Similarly, (7.1a) is satisfied since by definition π_y^w is invariant under the transition $P_{yv}^w = \sum_a P_{yav} p_{a|y}^w$.

To prove (iii), let $z(y) := \sum_{a \in A(y)} z(y, a)$ and substitute (7.4) in (7.1a) to obtain

$$z(y) \geqq \sum_{v \in \mathbf{X}} z(v) P_{vy}^w. \tag{7.6}$$

Following Lemma 3.1 and using (7.1c) we obtain $z(y) = \pi^w(y) = \bar{f}_{x,w}(y)$. By (7.4) and by the fact that $\bar{f}_{x,w}(y, a) = \bar{f}_{x,w}(y) \cdot p_{a|y}^w$ we obtain $z(y, a) = \pi^w(y) p_{a|y}^w = \bar{f}_{x,w}(y, a)$. It then follows by Lemma 2.3 and (7.1b) that $D^k(w) \leqq V_k$, $1 \leqq k \leqq K$, and therefore w is feasible for COP.

Parts (ii) and (iv) follow from the fact established above that (7.1) and (7.4) define a one-to-one correspondence between the z's that are feasible to LP and the stationary policies w's that are feasible to COP. Moreover, under this correspondence, it follows from Lemma 2.3 that the value $C(z)$ of LP is equal to $C_x(w)$, which establishes the proof. □

8. Extensions. In this section we outline some applications of our methods to lesser-known optimization criteria, involving variance minimization.

8.1. Variability sensitive optimization. The variability sensitive optimization problem VSOP was studied in the finite case by Filar, Kallenberg, and Lee [16] and later by Bayal-Gursoy and Ross [7];

$$\text{Maximize } R_x(u) := \varliminf_{t \to \infty} \frac{1}{t} \sum_{s=1}^{t} E_u[r(c(X_s, A_s), C_x^t(u))], \tag{8.1}$$

where $r(\cdot,\cdot)$ is called the variability function. Taking $r(x,y)=x-\lambda(x-y)^2$ the VSOP obtains the interpretation of finding a policy u that has high expected average reward but low expected variance. Other variability criteria and other variability functions are treated in the finite state-action space in [7] and [16].

In Theorem 8.1 we present conditions that ensure the sufficiency of classes of policies for problem VSOP. We will use $r(x,y)=x-\lambda(x-y)^2$. Note that when $\lambda=0$ this reduces to problem OP.

THEOREM 8.1. *Consider problem* VSOP. *Assume* (A1) *and* (A2) *and let U' be complete. If $\{c^2(X_s, A_s)\}_s$ is uniformly integrable with respect to P^u for each u, then U' is sufficient.*

Proof. First note that $R_x(u)$ is equal to

$$\underline{\lim}_{t\to\infty}\left[\sum_{y,a}\bar{f}^{\,t}_{x,u}(y,a)[c(y,a)-\lambda c^2(y,a)]+\lambda\left[\sum_{y,a}\bar{f}^{\,t}_{x,u}(y,a)c(y,a)\right]^2\right]. \tag{8.2}$$

Let t_n be any subsequence of t that achieves the $\underline{\lim}$ in the expression above, and along which $\bar{f}^{\,t_n}_{x,u}(y,a)\to\bar{f}_{x,u}(y,a)$ for all y and a. Following the same weak convergence arguments that were used in § 2.1, we obtain from the uniform integrability

$$R_x(u)=\sum_{y,a}\bar{f}_{x,u}(y,a)[c(y,a)-\lambda c^2(y,a)]+\lambda\left[\sum_{y,a}\bar{f}_{x,u}(y,a)c(y,a)\right]^2. \tag{8.3}$$

Thus $R_x(u)$ can be represented as a function of the expected state-action frequency, so completeness implies sufficiency. □

As a simple corollary, for bounded cost completeness implies sufficiency.

8.2. The problem with constraints. VSOP can also be considered in the framework of optimization under constraints. Kawai [23] introduced the problem of minimizing the variance of some cost subject to a single constraint on the expected average cost. He treats the case of finite state and action spaces, and restricting to the stationary policies he finds an optimal solution. Kurano [24] finds a policy that is optimal among the stationary deterministic policies for the same problem as Kawai yet with general state and action spaces.

Using similar arguments as above, we show below that any complete family of policies (e.g., the stationary policies) is sufficient for the problem of Kawai; hence the solution that Kawai finds is overall optimal. Moreover, using the same kind of assumptions as in Theorem 8.1 we show (using arguments as in the proof of Theorem 2.8) that these are sufficient for the case of countable state and action spaces, and for more than one constraint on expected average cost functionals.

Denote the variance under a policy u with initial state x by $R_x(u)$ through (8.1) with $r(x,y):=(x-y)^2$. Given K real numbers $V_1,\cdots,V_K$, define the following constrained problem:

$$\text{(CVSOP)}\qquad \text{minimize}\quad R_x(u)$$
$$\text{subject to}\quad \bar{D}^k_x(u)\leqq V_k,\qquad 1\leqq k\leqq K.$$

References [23] and [24] consider the case $V=V_1$ that is (ε) close to the supremum of the optimal expected average cost. The meaning of CVSOP is then to find a policy that minimizes the variance among all policies that are ε-optimal for OP.

THEOREM 8.2. *Consider problem* CVSOP. *Assume* (A1) *and* (A2) *and let U' be complete. If $\{c^2(X_s, A_s)\}_s$ and $\{d^k(X_s, A_s)\}_s$ $1\leqq k\leqq K$ are uniformly integrable with respect to P^u for each u, then U' is sufficient.*

Proof. The variance is given by

$$R_x(u)=\overline{\lim_{t\to\infty}}\left(\left[\sum_{y,a}\bar{f}^{t}_{x,u}(y,a)c(y,a)\right]^2-\sum_{y,a}\bar{f}^{t}_{x,u}(y,a)c^2(y,a)\right).$$

By diagonalization, there exists some subsequence t_n along which $\lim_{n\to\infty}\bar{f}^{t_n}_{x,u}(y,a)=\bar{f}_{x,u}(y,a)$ for all y and a, such that

$$R_x(u)=\left[\sum_{y,a}\bar{f}_{x,u}(y,a)c(y,a)\right]^2-\sum_{y,a}\bar{f}_{x,u}(y,a)c^2(y,a).$$

(similarly to the way (8.3) is obtained).

The rest of the proof now follows the same lines as the proof of Theorem 2.8. □

Acknowledgements. The authors wish to thank F. Spieksma and A. Hordijk for extensive discussions concerning this paper, and an anonymous reviewer for many constructive comments, which led to significant improvements in many respects.

The support and hospitality of the Mathematics of Networks and Systems Department, AT&T Bell Laboratories, Murray Hill, NJ, is gratefully acknowledged.

REFERENCES

[1] E. Altman and A. Shwartz, *Optimal priority assignment with general constraints*, in Proc. 24th Allerton Conference, University of Illinois, Urbana-Champaign, IL, October 1986.

[2] ———, *Non-stationary policies for controlled Markov chains*, EE Pub. 633, Technion-Israel Institute of Technology, Haifa, Israel, June 1987, submitted.

[3] ———, *Optimal priority assignment: a time sharing approach*, IEEE Trans. Automat. Control, 34 (1989), pp. 1098-1102.

[4] ———, *Adaptive control of constrained Markov chains*, IEEE Trans. Automat. Control, to appear (1991).

[5] J. S. Baras, A. J. Dorsey, and A. M. Makowski, *Discrete time competing queues with geometric service requirements: stability, parameter estimation and adaptive control*, SIAM J. Control Optim., submitted.

[6] J. S. Baras, D.-J. Ma, and A. M. Makowski, *K competing queues with geometric service requirements and linear costs: the μc rule is always optimal*, Systems Control Lett., 6 (1985), pp. 173-180.

[7] M. Bayal-Gursoy and K. W. Ross, *Variability sensitive Markov decision processes*, Math. Oper. Res., to appear.

[8] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[9] V. Borkar, *On minimum cost per unit time control of Markov Chains*, SIAM J. Control Optim., 22 (1983), pp. 965-978.

[10] ———, *Control of Markov chains with long-run average cost criterion*, in Proc. Stochastic Differential Systems, W. Fleming and P.-L. Lions, eds., Springer-Verlag, Berlin, New York, 1986, pp. 57-77.

[11] C. Buyyukkoc, P. Varaiya, and J. Walrand, *The $c\mu$ rule revisited*, Adv. Appl. Probab., 17 (1985), pp. 237-238.

[12] R. Cavazos-Cadena, *Existence of optimal stationary policies in average-reward Markov decision processes with a recurrent state*, Appl. Math. Optim., submitted.

[13] K. L. Chung, *Markov Chains with Stationary Transition Probabilities*, 2nd ed., Springer-Verlag, New York, 1967.

[14] R. Dekker and A. Hordijk, *Average, sensitive and Blackwell optimal policies in denumerable Markov decision chains with unbounded rewards*, Math. Oper. Res., 13 (1988), pp. 395-421.

[15] C. Derman, *Finite State Markovian Decision Processes*, Academic Press, New York, 1970.

[16] J. A. Filar, L. C. M. Kalllenberg, and H. M. Lee, *Variance penalized Markov decision processes*, Math. Oper. Res., 14 (1989), pp. 147-161.

[17] L. Fisher and S. M. Ross, *An example in denumerable decision processes*, Ann. Math. Statist., 39 (1968), pp. 674-675.

[18] L. Fisher, *On recurrent denumerable decision processes*, Ann. Math. Statist., 39 (1968), 424-434.

[19] E. Gelenbe and I. Mitrani, *Analysis and Synthesis of Computer Systems*, Academic Press, London, 1980.

[20] P. Hall and C. C. Heyde, *Martingale Limit Theory and Its Applications*, John Wiley, New York, 1980.

[21] A. Hordijk, *Dynamic Programming and Markov Potential Theory*, Mathematical Center Tracts, no. 51, Amsterdam, the Netherlands, 1974.

[22] A. Hordijk and L. C. M. Kallenberg, *Constrained undiscounted stochastic dynamic programming*, Math. Oper. Res., 9 (1984), pp. 276-289.

[23] H. Kawai, *A variance minimization problem for a Markov Decision process*, European J. Oper. Res., 31 (1987), pp. 140-145.

[24] M. Kurano, *Markov decision processes with a minimum-variance criterion*, J. Optim. Theory Anal., 123 (1987), pp. 572-583.

[25] A. M. Makowski and A. Shwartz, *Recurrence properties of a system of competing queues, with applications*, EE Pub. 627, Technion-Israel Institute of Technology, Haifa, Israel, 1987.

[26] P. Nain and K. W. Ross, *Optimal priority assignment with hard constraint*, IEEE Trans. Automat. Control, 31 (1986), pp. 883-888.

[27] D. Revuz, *Markov Chains*, North-Holland, Amsterdam, the Netherlands, 1975.

[28] L. I. Sennott, *A new condition for the existence of optimal stationary policies in average cost Markov decision processes*, Oper. Res. Lett., 5 (1986), pp. 17-23.

[29] ———, *Average cost optimal stationary policies in infinite state Markov decision processes with unbounded costs*, Oper. Res., 37 (1989), pp. 626-633.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 810-828, July 1991

STABILIZABILITY OF TIME-PERIODIC PARABOLIC EQUATIONS*

ALESSANDRA LUNARDI†

Abstract. Second-order parabolic equations with time-periodic coefficients in $[0, +\infty[\times \Omega$($\Omega$ is a regular bounded domain in $\mathbf{R}^n$) are considered, with input acting either in the interior or on the boundary of Ω. Stabilizability results that are similar to some of the known results concerning the autonomous case are given. Abstract evolution equations techniques are employed.

Key words. asymptotic behavior, stabilizability, evolution operator in periodic parabolic problems

AMS(MOS) subject classifications. 35K20, 35B35, 34G10, 49E20, 93D20, 93D99

1. Introduction. We consider nonautonomous parabolic problems defined on a bounded domain Ω in $\mathbf{R}^n$ with regular boundary $\partial\Omega$ and exterior unit normal vector $\nu(x) = (\nu_1(x), \cdots, \nu_n(x))$:

$$\begin{aligned} &u_t(t,x) = \mathscr{A}(t,x,\partial)u(t,x) + \phi(t,x), \qquad t>0, \quad x\in\Omega,\\ &u(0,x) = u_0(x), \qquad x\in\Omega,\\ &\mathscr{B}(x,\partial)u(t,x) = 0, \qquad t>0, \quad x\in\partial\Omega, \end{aligned} \tag{1.1}$$

$$\begin{aligned} &v_t(t,x) = \mathscr{A}(t,x,\partial)v(t,x), \qquad t>0, \quad x\in\Omega,\\ &v(0,x) = v_0(x), \qquad x\in\Omega,\\ &\mathscr{B}(x,\partial)v(t,x) = \psi(t,x), \qquad t>0, \quad x\in\partial\Omega. \end{aligned} \tag{1.2}$$

Here $\mathscr{A}(t,x,\partial) = a_{ij}(t,x)D_{ij} + b_i(t,x)D_i + c(t,x)\mathbf{I}$ is any elliptic operator with smooth coefficients, T-periodic with respect to time, and $\mathscr{B}(x,\partial)$ is either the trace operator on $\partial\Omega$ or any mixed nontangential boundary operator with smooth coefficients: $\mathscr{B}(x,\partial)u = \beta_i(\cdot)D_iu + \gamma(\cdot)u$ (the summation convention is used throughout). The inputs ϕ, ψ are of the form

$$\phi(t,x) = (\Phi(t)f(t))(x), \qquad \psi(t,x) = (\Psi(t)g(t))(x), \tag{1.3}$$

where f, g have values in general Banach spaces Y, Z, respectively, and $\Phi(t)$, $\Psi(t)$ are linear T-periodic operators belonging to $L(Y, L^p(\Omega))$, $L(Z, L^p(\partial\Omega))$, respectively.

The stabilizability problem consists in finding sufficient conditions on the data in order that for each u_0, $v_0 \in L^p(\Omega)$ there are $f:[0,+\infty[\to Y$, $g:[0,+\infty[\to Z$ such that the solutions u, v of problems (1.1), (1.2) decay asymptotically to zero as $t\to+\infty$, in the strongest possible norm. Obviously, for such a problem to be significant, we assume that the free systems (with $f\equiv 0$, $g\equiv 0$) are not asymptotically stable.

Systems of the type (1.1), (1.2) arise in several fields, like heat conduction and diffusion problems (see, e.g., [23, Chaps. 3, 5]).

We work in a L^p space setting ($1<p<+\infty$) to avoid technical difficulties. However, since we employ an abstract evolution equations approach, other topologies could, in principle, be considered. In fact, our analysis is based on the representation formulas

* Received by the editors December 27, 1988; accepted for publication (in revised form) August 10, 1990. The work was partially supported by the National Project Ministero della Pubblica Istruzione "Equazioni di Evoluzione e Applicazioni Fisico-Matematiche."

† Dipartimento di Matematica, Via Ospedale 72, 09124 Cagliari, Italy. The author is a member of Gruppo Nazionale Analisi Funzionale e Applicazioni of Consiglio Nazionale delle Ricerche.

for the solutions of (1.1), (1.2):

$$U(t,\cdot)=G(t,0)u_0+\int_0^t G(t,s)\phi(s,\cdot)\,ds,\qquad t\geqq 0, \tag{1.4}$$

$$v(t,\cdot)=G(t,0)v_0+\int_0^t G_s(t,s)H(s)\psi(s,\cdot)\,ds,\qquad t\geqq 0, \tag{1.5}$$

where $G(t,s)\in L(L^p(\Omega), W^{2,p}(\Omega))$ is the evolution operator of the free system, $G_s(t,s)=\partial/\partial s\; G(t,s)$ and $H(s)$ is either the Dirichlet mapping $D(s)$ or the mixed mapping $M(s)$ relevant to the elliptic operator $\mathcal{A}(s,\cdot,\partial)$, i.e., $H(s)g$ is the solution z of the problem

$$\mathcal{A}(s,\cdot,\partial)z\equiv 0\quad\text{in }\Omega,\qquad \mathcal{B}(\cdot,\partial)z=g\quad\text{in }\partial\Omega \tag{1.6}$$

(see [17] for formula (1.5)). Therefore, throughout the paper we consider a more general situation: we are given a couple of Banach spaces X, D (D being continuously embedded in X), and a family of generators of analytic semigroups $A(t): D\to X$ with $A(t+T)=A(t)$, such that there exists an evolution operator $G(t,s)$ for the linear problem $u'(t)=A(t)u(t)$.

In § 2 we state some asymptotic behavior properties (in the X-norm and in the D-norm) of the functions u, v defined in (1.4), (1.5). Such properties are consequences of results in [16] (where a Floquet theory for abstract parabolic equations has been developed) and [17] (where differentiability of $G(t,s)$ with respect to s has been shown, together with formula (1.5)). Floquet theory deals with asymptotic behavior of the evolution operator: as in ordinary differential equations, an essential role is played by the spectra of the linear operators

$$V(s)=G(s+T,s),\qquad s\in\mathbf{R}. \tag{1.7}$$

In particular, in our examples (1.1), (1.2), the free systems are not asymptotically stable if and only if some eigenvalues of $V(s)$ have modulus greater or equal to one (we recall that the nonzero eigenvalues of $V(s)$ are independent of s, whereas the corresponding eigenvectors may depend explicitly on s).

In § 3 we study stabilizability of abstract parabolic equations:

$$u'(t)=A(t)u(t)+\Phi(t)f(t),\qquad t>0;\quad u(0)=u_0, \tag{1.8}$$

where $\Phi(t)\in L(Y,X)$ and $\Phi(t+T)=\Phi(t)$, Y being any Banach space. We find a stabilizability result that generalizes the well-known results concerning ordinary differential equations (see [11]). Precisely, we fix $\omega\geqq 0$ such that no element of the spectrum of $V(s)$ has modulus $e^{-\omega T}$, and we show that the following conditions are equivalent:

(i) For every u_0 in the closure of D there is a continuous $f:[0,+\infty[\to Y$, with $\|e^{\omega t}f(t)\|_Y$ bounded, such that $\limsup_{t\to+\infty}\|e^{\omega t}u(t)\|_D<+\infty$.

(ii) For every $\lambda\in\mathbf{C}$ with $|\lambda|>e^{-\omega T}$ we have (denoting, as usual, adjoint operators by *)

$$(\lambda-V(0)^*)x^*=0,\quad s\to\Phi(s)^*G(T,s)^*x^*\equiv 0\quad\text{in }[0,T]\Rightarrow x^*=0.$$

In § 3.2 we apply the result of (1.1), with

$$Y=\mathbf{R}^k,\qquad \Phi(t)(y_1,\cdots,y_k)=\phi_i(t,\cdot)y_i, \tag{1.9}$$

where $t\to\phi_i(t,\cdot)$ belongs to $C^\alpha(\mathbf{R};L^p(\Omega))$, $\phi_i(t+T,\cdot)=\phi_i(t,\cdot)$, $i=1,\cdots,k$, and $\phi_1,\cdots,\phi_k$ are linearly independent. We find that system (1.1) is stabilizable in the

$W^{2,p}$-norm if an algebraic condition, involving the functions ϕ_i and the eigenfunctions of $V(s)^*$, holds. Precisely, we require that for each eigenvalue λ of $V(0)$ with $|\lambda| \geqq 1$ there is $s \in [0, T[$ such that, denoting by $\{\xi_1^*, \cdots, \xi_{N(\lambda)}^*\}$ any system of generators of $\ker(\lambda - V(s)^*)$, we have $k \geqq N(\lambda)$, and the matrix

$$[A_{ji}(s)] = [\langle \phi_i(s, \cdot), \xi_j^* \rangle]_{i=1,\cdots,k; j=1,\cdots,N(\lambda)} \tag{1.10}$$

has rank $N(\lambda)$ (we identify, as usual, $(L^p(\Omega))^*$ and $L^{p^*}(\Omega)(1/p + 1/p^* = 1)$, setting $\langle f, g^* \rangle = \int_\Omega f(x) g^*(x)\, dx$). In the particular case of the one-dimensional heat equation with time-periodic coefficients (see [23, Chap. 5]):

$$\begin{aligned} &u_t(t, x) = a(t) u_{xx}(t, x) + c(t) u(t, x) + y(t)\phi(x), \qquad 0 < x < l, \quad t > 0, \\ &u(0, x) = u_0(x), \qquad 0 < x < l, \\ &u(t, 0) = u(t, l) = 0, \qquad t > 0, \end{aligned} \tag{1.11}$$

with $a(t)$, $c(t) > 0$, $\phi \in L^p(0, l)$, the evolution operator is $G(t, s) = \exp[A \int_s^t a(s)\, ds + \int_s^t c(s)\, ds]$ (where e^{At} is the semigroup generated by the second derivative operator with zero boundary condition), and the stabilizability condition is the following: For each $k \in \mathbf{N}$ such that $\int_0^T c(s)\, ds - \pi^2 k^2 l^{-2} \int_0^T a(s)\, ds > 0$, we have $\int_0^l \phi(x) \sin(\pi k x / l)\, dx \neq 0$.

In § 4 we consider boundary stabilizability. Formally, the results are similar to the previous ones. In the case of Dirichlet boundary condition (§ 4.1) we fix $\omega \geqq 0$ such that no eigenvalue of $V(0)$ has modulus $e^{-\omega T}$, and we show that the following conditions are equivalent:

(i) For every $u_0 \in L^p(\Omega)$ there is a regular function $g : [0, +\infty[\to Z$ (with $\|e^{\omega t} g(t)\|_Z$ bounded) such that $\limsup_{t \to +\infty} \|e^{\omega t} u(t)\|_{W^{2,p}(\Omega)} < +\infty$,

(ii) For every $\lambda \in \mathbf{C}$ with $|\lambda| > e^{-\omega T}$ we have

$$(\lambda - V(0)^*)x^* = 0, \quad s \to \Psi(s)^* \left[\frac{\partial}{\partial \nu_s} (G(T, s)^* x^*) \right] \equiv 0 \quad \text{in } [0, T] \Rightarrow x^* = 0.$$

Here $\partial/\partial \nu_s$ denotes the conormal derivative (at the boundary) at the time s: $\partial/\partial \nu_s f = a_{ij}(s, \cdot) D_i f(\cdot) \nu_j(\cdot)$.

In particular, in the case where $Z = W^{2-1/p,p}(\partial\Omega)$, $\Psi(s) \equiv 1$, condition (ii) is fulfilled (at least in the self-adjoint, C^∞ case) thanks to a result in [19]. In the case where $Z = \mathbf{R}^k$, $\Psi(t)(z_1, \cdots, z_k) = \psi_i(t, \cdot) z_i$ (where $\psi_i(t, \cdot)$ are regular, T-periodic, linearly independent functions with values in $W^{2-1/p,p}(\partial\Omega)$), (ii) is fulfilled provided a full rank condition holds, i.e., there is $s \in \mathbf{R}$ such that $k \geqq N(\lambda)$ and the matrix

$$[B_{ij}] = \left[\left\langle \psi_i(s, \cdot), \frac{\partial}{\partial \nu_s} \xi_j^* \right\rangle \right]_{i=1,\cdots,k; j=1,\cdots,N(\lambda)} \tag{1.12}$$

has rank $N(\lambda)$ (again, we identify $(L^p(\partial\Omega))^*$ and $L^{p^*}(\partial\Omega)$, setting $\langle f, g^* \rangle = \int_{\partial\Omega} f(x) g^*(x)\, d\sigma_x$, and, as before, $\{\xi_1^*, \cdots, \xi_{N(\lambda)}^*\}$ is any system of generators of $\ker(\lambda - V(s)^*)$).

In the case of mixed boundary condition (§ 4.2) we find quite similar results, with $\partial/\partial \nu_s$ replaced by the trace operator. For instance, in the case of the time-periodic heat equation with Neumann boundary condition,

$$\begin{aligned} &u_t(t, x) = a(t) \Delta u(t, x), \qquad x \in \Omega, \quad t > 0, \\ &u(0, x) = u_0(x), \qquad x \in \Omega, \\ &\frac{\partial}{\partial \nu} u(t, x) = z(t) \psi(t, x), \qquad t > 0, \end{aligned} \tag{1.13}$$

the unique eigenvalue of $G(T, 0)$ with modulus greater than or equal to one is one and the corresponding eigenspace is one-dimensional (it consists of the constant functions), so that the system is exponentially stabilizable provided there is s such that $\int_{\partial\Omega} \psi(s, x)\, d\sigma_x \neq 0$.

Concerning the literature on the subject, a considerable amount of papers have been devoted to stabilizability (in fact, feedback stabilizability) in autonomous problems, both in the case of distributed and boundary inputs (see, for instance, [5], [18], [21], [12], [13], [4], and the references quoted there). To the author's knowledge, stabilizability in nonautonomous parabolic problems has not yet been studied.

We remark that abstract and concrete feedback stabilizability results could be obtained from ours (at least, in the case where X, Y are Hilbert spaces, i.e., $p=2$) using the theories developed in [7] and [8]. Concerning problem (1.8), it has been shown in [7] that if the system is stabilizable then there is an optimal pair (u_0^*, f^*) minimizing the cost functional

$$J(u_0, f) = \int_0^{+\infty} (\|u(t)\|_X^2 + \|f(t)\|_Y^2)\, dt, \tag{1.14}$$

and, denoting by u^* the solution of (1.8) with data (u_0^*, f^*) we have $f^*(t) = -\Phi(t)^* Q(t) u^*(t)$, where $Q(t) \in L(X)$ is the unique positive T-periodic solution of the Riccati equation

$$Q'(t) + A(t)^* Q(t) + Q(t)A(t) - Q(t)\Phi(t)\Phi(t)^* Q(t) + \mathbf{I} = 0, \qquad t \in \mathbf{R}. \tag{1.15}$$

From [6] it follows that the evolution operator generated by $A - \Phi\Phi^* Q$ is exponentially stable (see § 3 of [7]): therefore we get the feedback $F(t) = -\Phi(t)^* Q(t)$.

A similar procedure may be employed in problem (1.2): in this case, the study of the corresponding Riccati equation is much more complicated. Solvability of initial value problems for such equations was shown in [1], where the existing results for the autonomous case ([13], [9]) were extended to a larger class of nonautonomous equations. The result of [1] also implies, as remarked in [8], the existence of a unique positive T-periodic solution of the Riccati equation. Arguing as before, we can construct a feedback operator.

2. Asymptotic behavior in abstract parabolic Cauchy problems.

Let X, D be Banach spaces, with norms $\|\cdot\|$, $\|\cdot\|_D$, respectively, D being continuously embedded in X. We consider a time-periodic initial value problem:

$$u'(t) = A(t)u(t) + f(t), \qquad t > 0, \quad u(0) = x. \tag{2.1}$$

The linear operators $A(t)$ are subject to the following assumptions ($0 < \alpha < 1$, $T > 0$):

(2.2)
(i) $t \to A(t) \in C^\alpha(\mathbf{R}; L(D, X))$,
(ii) $A(t+T) = A(t)$, $\quad t \in \mathbf{R}$,
(iii) For each $t \in \mathbf{R}$, $A(t)$ generates an analytic semigroup $e^{sA(t)}$ in X and there is $c \geqq 1$ such that $c^{-1}\|x\|_D \leqq \|x\| + \|A(t)x\| \leqq c\|x\|_D$ for every $x \in D$, $t \in \mathbf{R}$.

Then there is an evolution operator $G(t, s) \in L(X, D)(t \geqq s)$, such that for every locally α-Hölder continuous function $f: [0, +\infty[\to X$ and for every x in the closure of D, problem (2.1) has a unique solution $u \in C[0, +\infty[; X) \cap C^1(]0, +\infty[; X) \cap$

$C(]0, +\infty[; D)$, given by the variation of constants formula:

$$u(t) = G(t, 0)x + \int_0^t G(t, s)f(s)\, ds, \qquad t \geqq 0. \tag{2.3}$$

If f is continuous in $[0, +\infty[$, then the function u defined in (2.3) is the unique strong solution of (2.1). For this and other properties of $G(t, s)$ we refer to [20] for the dense domain case, and to [15] for the nondense domain case.

Some regularity properties will be described through the interpolation spaces X_θ $(0 < \theta < 1)$ defined by

$$X_\theta = (X, D)_{\theta,\infty} \tag{2.4}$$

(see, e.g., [22]). To treat asymptotic behavior as $t \to +\infty$, we introduce certain functional spaces consisting of exponentially decaying functions: for any Banach space B, $\omega \geqq 0$ and $0 < \alpha < 1$, we set

$$\begin{aligned} &C_\omega([0, +\infty[; B) = \{u \in C([0, +\infty[; B); \sup_{t>0} \|u(t)\, e^{\omega t}\|_B < +\infty\}, \\ &\|u\|_{C_\omega([0,+\infty[; B)} = \sup_{t>0} \|u(t)\, e^{\omega t}\|_B, \end{aligned} \tag{2.5}$$

$$\begin{aligned} &C^\alpha_\omega([0, +\infty[; B) \\ &\qquad = \left\{u \in C_\omega([0, +\infty[; B); \sup_{0<s<t} \|u(t)\, e^{\omega t} - u(s)\, e^{\omega s}\|_B (t-s)^{-\alpha} < +\infty\right\}, \\ &\|u\|_{C^\alpha_\omega([0,+\infty[; B)} \\ &\qquad = \sup_{t>0} \|u(t)\, e^{\omega t}\|_B + \sup_{t>s>0} \|u(t)\, e^{\omega t} - u(s)\, e^{\omega s}\|_B (t-s)^{-\alpha}, \end{aligned} \tag{2.6}$$

$$\begin{aligned} &C^1_\omega([0, +\infty[; B) = \{u \in C^1([0, +\infty[; B); u, u' \in C_\omega([0, +\infty[; B))\}; \\ &\|u\|_{C^1_\omega([0,+\infty[; B)} = \|u\|_{C_\omega([0,+\infty[; B)} + \|u'\|_{C_\omega([0,+\infty[; B)}. \end{aligned} \tag{2.7}$$

In the limiting case $\omega = 0$, $C_0([0, +\infty[; B)$ consists of bounded functions, and $C^\alpha_0([0, +\infty[; B)$ consists of uniformly Hölder continuous functions. It is easy to see that if u belongs to $C^\alpha_\omega([0, +\infty[; B)$, then

$$\sup_{0<s<t} e^{\omega s} \|u(t) - u(s)\|_B (t-s)^{-\alpha} \leqq \gamma(\omega) \|u\|_{C^\alpha_\omega([0,+\infty[; B)}. \tag{2.8}$$

We now recall some results of [16], which generalize to the infinite-dimensional case the known asymptotic behavior properties of the evolution operator for ordinary differential equations. We define the period map (Poincaré map) by

$$V(s) = G(s + T, s), \qquad s \in \mathbf{R}. \tag{2.9}$$

Denote by $\sigma(V(s))$ the spectrum of $V(s)$, and fix $\omega \geqq 0$ such that

$$\sigma(V(s)) \cap \{\lambda \in \mathbf{C}; |\lambda| = e^{-\omega T}\} = \varnothing, \qquad s \in \mathbf{R}. \tag{2.10}$$

Since $s \to V(s)$ is uniformly continuous (in fact, α-Hölder continuous due to [15, Prop. 3.6]) with values in $L(X)$ and $\sigma(V(s))$ is closed for every s, then, setting

$$\begin{aligned} &\sigma(V(s)) = \sigma_1(V(s)) \cup \sigma_2(V(s)), \qquad s \in \mathbf{R}, \\ &\sigma_1(V(s)) \subset \{\lambda \in \mathbf{C}; |\lambda| < e^{-\omega T}\}, \quad \sigma_2(V(s)) \subset \{\lambda \in \mathbf{C}; |\lambda| > e^{-\omega T}\} \quad \forall s \in \mathbf{R}, \end{aligned} \tag{2.11}$$

we have

$$(2.12)\qquad \rho_1=\sup\{|\lambda|;\ \lambda\in\sigma_1(V(s)),\ s\in\mathbf{R}\}<e^{-\omega T}<\rho_2=\inf\{|\lambda|;\ \lambda\in\sigma_2(V(s)),\ s\in\mathbf{R}\}.$$

Two families of projections may be defined as follows:

$$(2.13)\qquad P_1(t)=\frac{1}{2\pi i}\int_{C(0,e^{-\omega T})}(\lambda-V(t))^{-1}\,d\lambda,\quad P_2(t)=(\mathbf{I}-P_1(t)),\quad t\in\mathbf{R},$$

where $C(0,r)$ is the circle centered at $0\in\mathbf{C}$ with radius $r>0$. The restriction of the evolution operator $G(t,s)$ to the subspace $P_2(s)(X)$ is well defined also for $t<s$, and we have, for every $\varepsilon>0$ sufficiently small ($\varepsilon<T^{-1}\log(e^{-\omega T}/\rho_1)$, $\varepsilon<T^{-1}\log(\rho_2/e^{-\omega T})$):

$$(2.14)\qquad \begin{aligned}&\text{(i)}\quad \|G(t,s)P_1(s)\|_{L(X)}\leqq k_1(\varepsilon)\,e^{-(\omega+\varepsilon)(t-s)},\quad t>s,\\ &\text{(ii)}\quad \|G(t,s)P_1(s)\|_{L(X,D)}\leqq k_1(\varepsilon)(t-s)^{-1}\,e^{-(\omega+\varepsilon)(t-s)},\quad t>s,\\ &\text{(iii)}\quad \|G(t,s)P_2(s)\|_{L(X,D)}\leqq k_2(\varepsilon)\,e^{(\omega+\varepsilon)(t-s)},\quad t<s\end{aligned}$$

(see [16, Props. 2.3(i), 2.7(i)]).

PROPOSITION 2.1. *Let* (2.2), (2.10) *hold, let f belong to $C_\omega([0,+\infty[;X)$, and let x belong to the closure of D in X. Then the function u defined in* (2.3) *belongs to $C_\omega([0,+\infty[;X)$ if and only if*

$$(2.15)\qquad P_2(0)x=-\int_0^{+\infty}G(0,s)P_2(s)f(s)\,ds.$$

If (2.15) *holds, and, in addition, f belongs to $C_\omega^\alpha([0,+\infty[;X)$, then u belongs to $C_\omega([a,+\infty[;D)$ for every $a>0$; in particular,*

$$(2.16)\qquad \sup_{t\geqq a}\|u(t)\,e^{\omega t}\|_D<+\infty\quad \textit{for each } a>0.$$

If (2.15) *holds, and, in addition, f belongs to $C_\omega([0,+\infty[;X_\theta)$ for some $\theta\in\,]0,1[$, then u' and $A(\cdot)u(\cdot)$ belong to $C_\omega([a,+\infty[;X)$ for every $a>0$; in particular* (2.16) *holds.*

Proof. The first part of the proposition was proved in Proposition 3.10 of [16]. Moreover, if f belongs to $C_\omega^\alpha([0,+\infty[;X)$, in particular, it belongs to $C^\alpha([0,a];X)$ for every $a>0$, so that, by Propostions 3.5(ii) and 2.6(i) of [15], u belongs to $C^\alpha([a/2,a];D)\cap C^{\alpha+1}([a/2,a];X)$. Consequently, by Lemma 1.1 of [15], $u'(a)=A(a)u(a)+f(a)$ belongs to X_α. Analogously, if f belongs to $C_\omega([0,+\infty[;X_\theta)$, then it belongs to $C([0,a];X_\theta)$ for every $a>0$, so that, by Propositions 2.5 and 3.5(iii) of [15], $A(t)u(t)$ belongs to X_θ for $0<t\leqq a$; in particular, $A(a)u(a)$ belongs to X_θ. Moreover, condition (2.15) implies easily

$$(2.17)\qquad P_2(a)u(a)=-\int_a^{+\infty}G(a,s)P_2(s)f(s)\,ds,$$

so that we can apply Proposition 3.10 of [16] to show the last part of the statement. □

Under some more regularity assumptions on $A(\cdot)$, i.e.,

$$(2.18)\qquad t\to A(t)\in C^{1+\alpha}(\mathbf{R},L(D,X))\quad\text{for some }\alpha\in\,]0,1[,$$

we can show that $s\to G(t,s)$ is differentiable for $t>s$ with values in $L(X,D)$. In the sequel, together with the function u defined in (2.3), we will be concerned also with the function

$$(2.19)\qquad v(t)=G(t,0)x+\int_0^t G_s(t,s)f(s)\,ds,\qquad t\geqq 0,$$

where f has values in some interpolation space X_θ, $0<\theta<1$. v is well defined, since (see [17]) for every $a>0$ there is $M_a>0$ such that

$$\|G_s(t,s)\|_{L(X_\theta,X)}\leqq M_a(t-s)^{\theta-1},\qquad 0\leqq s<t\leqq a. \tag{2.20}$$

In order to deal with asymptotic behavior of $v(t)$, we establish some estimates for $P_i(t)G_s(t,s)$, $i=1,2$.

LEMMA 2.2. *Let* (2.2), (2.10), (2.18) *hold. Then*:

(i) *For every* $t\in\mathbf{R}$, *the function* $s\to G(t,s)P_1(s)$ *is differentiable in* $]-\infty,t[$ *with values in* $L(X,D)$, *and*

$$\frac{\partial}{\partial s}(G(t,s)P_1(s))=P_1(t)G_s(t,s),\qquad t>s. \tag{2.21}$$

Moreover, for every $\varepsilon\in[0,T^{-1}\log(e^{-\omega T}/\rho_1)[$ *there is* $C_1(\varepsilon)>0$ *such that*

$$\begin{aligned}
&\text{(i)}\quad \left\|\frac{\partial}{\partial s}(G(t,s)P_1(s))x\right\|\\
&\qquad\leqq C_1(\varepsilon)(t-s)^{\theta-1}\,e^{-(\omega+\varepsilon)(t-s)}\|x\|_\theta,\qquad t>s,\quad x\in X_\theta,\\
&\text{(ii)}\quad \left\|\frac{\partial}{\partial s}(G(t,s)P_1(s))x\right\|_D\\
&\qquad\leqq C_1(\varepsilon)(t-s)^{\theta-2}\,e^{-(\omega+\varepsilon)(t-s)}\|x\|_\theta,\qquad t>s,\quad x\in X_\theta.
\end{aligned} \tag{2.22}$$

(ii) *For every* $t\in\mathbf{R}$, *the function* $s\to G(t,s)P_2(s)$ *is differentiable in* $\mathbf{R}$ *with values in* $L(X,D)$, *and*

$$\begin{aligned}
&\frac{\partial}{\partial s}(G(t,s)P_2(s))=P_2(t)G_s(t,s),\qquad t\geqq s,\\
&\frac{\partial}{\partial s}(G(t,s)P_2(s))=G(t,r)P_2(r)G_s(r,s)=G(t,r)G_s(r,s),\qquad t\leqq s<r.
\end{aligned} \tag{2.23}$$

Moreover, for every $\varepsilon\in[0,T^{-1}\log(\rho_2/e^{-\omega T})[$ *there is* $C_2(\varepsilon)>0$ *such that*

$$\left\|\frac{\partial}{\partial s}(G(t,s)P_2(s))x\right\|_D\leqq C_2(\varepsilon)\,e^{(\omega+\varepsilon)(t-s)}\|x\|_\theta,\qquad t\leqq s,\quad x\in X_\theta. \tag{2.24}$$

Proof. (i) Since $G(t,s)P_1(s)=P_1(t)G(t,s)$ for every $t>s$, we have, for $t>s$, $t>s+h$:

$$h^{-1}[G(t,s+h)P_1(s+h)-G(t,s)P_1(s)]=P_1(t)h^{-1}[G(t,s+h)-G(t,s)].$$

Therefore $G(t,\cdot)P_1(\cdot)$ is differentiable in $]-\infty,t[$ with values in $L(X)$, and (2.21) holds. Since $G(t,s)=G(t,r)G(r,s)$ for $t>r>s$, then $G_s(t,s)=G(t,r)G_s(r,s)$, so that $G(t,\cdot)P_1(\cdot)$ is differentiable in $]-\infty,t[$ with values in $L(X,D)$, and

$$P_1(t)G_s(t,s)=P_1(t)G(t,r)G_s(r,s)=G(t,r)P_1(r)G_s(r,s),\qquad t>r>s. \tag{2.25}$$

If $t\geqq s+2$ we get, choosing $r=s+1$ and using estimates (2.14) and (2.20) with $a=1$,

$$\left\|\frac{\partial}{\partial s}(G(t,s)P_1(s))x\right\|+\left\|\frac{\partial}{\partial s}(G(t,s)P_1(s))x\right\|_D\leqq 2M_1k_1(\varepsilon)\,e^{-(\omega+\varepsilon)(t-s)}\|x\|_\theta.$$

If $t\leqq s+2$ we choose $r=(t+s)/2$ in (2.25), and, again using (2.14) and (2.20) we find

$$\begin{aligned}
&\left\|\frac{\partial}{\partial s}(G(t,s)P_1(s))x\right\|\leqq k_1(\varepsilon)\,e^{-2(\omega+\varepsilon)}M_2\left[\frac{t-s}{2}\right]^{\theta-1}\|x\|_\theta,\\
&\left\|\frac{\partial}{\partial s}(G(t,s)P_1(s))x\right\|_D\leqq\left[\frac{t-s}{2}\right]^{-1}k_1(\varepsilon)\,e^{-2(\omega+\varepsilon)}M_2\left[\frac{t-s}{2}\right]^{\theta-1}\|x\|_\theta.
\end{aligned}$$

By the arbitrariness of ε we get estimates (2.22).

The proof of (ii) is analogous; to show (2.24) we can use the second equality in (2.23) with $r=1+s$, and estimate (2.14)(iii). □

Now we are ready to prove a result similar to Proposition 2.1.

PROPOSITION 2.3. *Let* (2.2), (2.10), (2.18) *hold, and let x belong to the closure of D, f belong to $C_\omega([0,+\infty[;X_\theta)$ for some $\theta\in\,]0,1[$. Then the function v defined in* (2.19) *belongs to $C_\omega([0,+\infty[;X)$ if and only if*

$$P_2(0)x=-\int_0^{+\infty}\frac{\partial}{\partial s}(G(0,s)P_2(s))f(s)\,ds. \tag{2.26}$$

If (2.26) *holds, and, in addition, f belongs to $C_\omega^\beta([0,+\infty[;X_\theta)$ with $\theta+\beta>1$, then v is differentiable for $t>0$, $v(t)-f(t)$ belongs to D for $t>0$, and*

$$v'(t)=A(t)[v(t)-f(t)],\qquad t>0. \tag{2.27}$$

Moreover, v', $A(\cdot)[v(\cdot)-f(\cdot)]$ belong to $C_\omega([a,+\infty[;X)$ for every $a>0$, and we have

$$\sup_{t\geqq a}\|(v(t)-f(t))\,e^{\omega t}\|_D\leqq C(a,\omega,\theta)(\|x\|+\|f\|_{C_\omega^\beta([0,+\infty[;X_\theta)}). \tag{2.28}$$

Proof. First of all, we note that, due to estimates (2.22) and (2.24), the functions

$$v_1(t)=P_1(t)v(t)=G(t,0)P_1(0)x+\int_0^t\left[\frac{\partial}{\partial s}(G(t,s)P_1(s))\right]f(s)\,ds,\qquad t>0,$$

$$v_2(t)=-\int_t^{+\infty}\left[\frac{\partial}{\partial s}(G(t,s)P_2(s))\right]f(s)\,ds,\qquad t\geqq 0,$$

belong both to $C_\omega([0,+\infty[;X)$ (in fact, v_2 belongs to $C_\omega([0,+\infty[;D))$.

For every $t\geqq 0$ we have $v(t)=P_1(t)v(t)+P_2(t)v(t)$. Hence, using (2.23) we get

$$\begin{aligned}v(t)-v_1(t)-v_2(t)&=G(t,0)P_2(0)x+\int_0^{+\infty}\left[\frac{\partial}{\partial s}(G(t,s)P_2(s))\right]f(s)\,ds\\&=G(t,0)\left\{P_2(0)x+\int_0^{+\infty}\left[\frac{\partial}{\partial s}(G(0,s)P_2(s))\right]f(s)\,ds\right\}\overset{\text{def}}{=}G(t,0)y,\end{aligned}$$

where y belongs to $P_2(0)(X)$. Since $\sup_{t>0}\|e^{\omega t}G(t,0)y\|=+\infty$ for every $y\in P_2(0)\times(X)\backslash\{0\}$, then v belongs to $C_\omega([0,+\infty[;X)$ if and only if $y=0$, i.e., (2.26) holds.

Let now (2.26) hold, and let f belong to $C_\omega^\beta([0,+\infty[;X_\theta)$, with $\theta+\beta>1$. For every $t>0$ we have, by Corollary 2.6 of [17] and equality (2.21),

$$\begin{aligned}&\int_0^t\frac{\partial}{\partial s}[G(t,s)P_1(s)]f(s)\,ds\\&\quad=\int_0^t P_1(t)G_s(t,s)[f(s)-f(t)]\,ds+P_1(t)\int_0^t G_s(t,s)f(t)\,ds\\&\quad=\int_0^t\frac{\partial}{\partial s}[G(t,s)P_1(s)][f(s)-f(t)]\,ds+P_1(t)[f(t)-G(t,0)f(t)],\qquad t\geqq 0,\end{aligned}$$

so that

$$v_1(t)=G(t,0)P_1(0)(x-f(t))+\int_0^t\frac{\partial}{\partial s}(G(t,s)P_1(s))[f(s)-f(t)]\,ds+P_1(t)f(t),\qquad t\geqq 0.$$

Therefore, by estimate (2.22)(ii), $v_1(t)-P_1(t)f(t)$ belongs to D for every $t>0$ and also using estimates (2.14)(ii) and (2.8) we get, for $t \geqq a > 0$,

$$\begin{aligned}\|[v_1(t) - P_1(t)f(t)]\, e^{\omega t}\|_D &\leqq t^{-1}k_1(0)\Big(\|x\| + \sup_{t>0} \|f(t)\|\Big) \\ &\quad + C_1(\varepsilon) \int_0^t \varepsilon^{-\varepsilon(t-s)}(\tau - s)^{\theta+\beta-2}\, ds \|f\|_{C^\beta_\omega([0,+\infty[;X_\theta)} \\ &\leqq a^{-1}k_1(0)\|x\| + [a^{-1}k_1(0) \\ &\quad + C_1(\varepsilon)\gamma(\beta,\omega)\varepsilon^{1-\theta-\beta}\Gamma(\theta+\beta-1)]\|f\|_{C^\beta_\omega([0,+\infty[;X_\theta)}\end{aligned}$$

so that $v_1 - P_1 f$ belongs to $C_\omega([0, +\infty[; D)$. Since v_2 also belongs to $C_\omega([0, +\infty[; D)$, and P_2 is bounded with values in $L(X, D)$ (see (2.13)), then $v_2 - P_2 f$ belongs to $C_\omega([0, +\infty[; D)$. Therefore $v - f = v_1 + v_2 - (P_1 + P_2)f$ belongs to $C_\omega([0, +\infty[; D)$. Moreover, by Remark 2.8 of [17], $v - G(\cdot, 0)x$ belongs to $C^1(]0, +\infty[; X)$, and $d/dt(v(t) - G(t, 0)x) = A(t)[v(t) - G(t, 0)x - f(t)]$ for $t > 0$. Since $G(\cdot, 0)x$ also belongs to $C^1(]0, +\infty[; X)$, then v belongs to $C^1(]0, +\infty[; X)$, and (2.27) holds. Formula (2.28) now follows easily. □

3. Stabilizability for distributed parameters systems. Here we apply the asymptotic behavior results of the previous section to prove some stabilizability theorems. We begin (§ 3.1) with an abstract stabilizability result generalizing the well-known one [11] concerning ordinary differential systems. We apply such a result to a distributed interior control problem (§ 3.2).

3.1. Stabilizability in abstract parabolic periodic problems. Throughout the section we use notation from § 2 and we assume that (2.2), (2.8) hold. We denote by Y another Banach space, with norm $\|\cdot\|_Y$, and consider the problem

$$u'(t) = A(t)u(t) + \Phi(t)f(t), \quad t > 0, \quad u(0) = x, \tag{3.1}$$

under the following assumptions on the linear operators $\Phi(t): Y \to X$ $(0 < \alpha < 1)$:

$$t \to \Phi(t) \in C^\alpha(\mathbf{R}; L(Y, X)), \quad \Phi(t + T) = \Phi(t) \quad \forall t \in \mathbf{R}. \tag{3.2}$$

For every x in the closure of D and $f \in C([0, +\infty[; Y)$, problem (3.1) has a unique strong solution u, given by the variation of constants formula,

$$u(t) = G(t, 0)x + \int_0^t G(t, s)\Phi(s)f(s)\, ds, \qquad t \geqq 0. \tag{3.3}$$

We want to find necessary and sufficient conditions on the family $\{\Phi(t); t \in \mathbf{R}\}$ in order that for every x belonging to the closure of D there is $f: [0, +\infty[\to Y$ such that the (strong) solution of (3.1) decays exponentially as $t \to +\infty$. Due to Proposition 2.1, the problem is not trivial if

$$\sigma(V(s)) \cap \{\lambda \in \mathbf{C}; |\lambda| \geqq 1\} \neq \emptyset \quad \text{for some } s \in \mathbf{R}. \tag{3.4}$$

We fix, once and for all, a number $\omega \geqq 0$ satisfying (2.10), and we assume (see (2.11)):

(3.5) For every $s \in \mathbf{R}$, $\sigma_2(V(s))$ consists of a finite number of eigenvalues with finite algebraic multiplicity.

It can be easily shown that the nonzero eigenvalues of $V(s)$ do not depend on s. Therefore we can set

$$\sigma_2 \overset{\text{def}}{=} \sigma_2(V(s)) \quad \forall s \in \mathbf{R}. \tag{3.6}$$

Again, it is easy to see that λ belongs to σ_2 if and only if the problem

$$v'(t)=A(t)v(t)+\kappa v(t), \qquad t\in\mathbf{R}, \tag{3.7}$$

with $\kappa=e^{-\lambda T}$ has a nontrivial T-periodic solution.

THEOREM 3.1. *Let* (2.2), (2.10), (3.3)-(3.5) *hold. The following conditions are equivalent*:

(i) *For every x in the closure of D there is $f\in C_\omega([0,+\infty[;\ Y)$ such that the solution u of* (3.1) *belongs to $C_\omega([0,+\infty[;\ X)$*;

(ii) *For every $\lambda\in\mathbf{C}$ with $|\lambda|>e^{-\omega T}$ we have*

$$(\lambda-V(0)^*)x^*=0,\quad \Phi(s)^*G(T,s)^*x^*=0 \quad \textit{for each } s\in[0,T]\Rightarrow x^*=0.$$

If one of the equivalent conditions (i) *or* (ii) *holds, then for each x belonging to the closure of D there is f belonging to $C^\alpha_\omega([0,+\infty[;\ Y)$, such that u belongs to $C_\omega([a,+\infty[;\ D)$ for every $a>0$.*

Proof. (i)$\Rightarrow$(ii) Assume by contradiction that (ii) does not hold. Choose $\lambda\in\mathbf{C}$ with $|\lambda|>e^{-\omega T}$ and $x^*\neq 0$ such that $\lambda x^*=V(0)^*x^*$, $\Phi(s)^*G(T,s)^*x^*\equiv 0$ in $[0,T]$. Let $x\in P_2(0)(X)$ be such that $\langle x,x^*\rangle=1$. By assumption there is $f\in C_\omega([0,+\infty[;\ Y)$ such that u belongs to $C_\omega([0,+\infty[;\ X)$. Taking $t=nT$, $n\in\mathbf{N}$, we have

$$\begin{aligned}\langle u(nT),x^*\rangle&=\langle G(nT,0)x,x^*\rangle+\int_0^{nT}\langle G(nT,s)\Phi(s)f(s),x^*\rangle\,ds\\&=\langle x,(V(0)^*)^nx^*\rangle\\&\quad+\sum_{k=0}^{n-1}\int_0^T\langle f(s+kT),\Phi(s)^*G(T,s)^*(V(0)^*)^{n-k-1}x^*\rangle\,ds\\&=\lambda^n\,e^{n\omega T}.\end{aligned}$$

Since $|\lambda^n e^{n\omega T}|=(|\lambda|\,e^{\omega T})^n\to+\infty$ as $n\to+\infty$, then $\sup_{t>0}\|u(t)\,e^{\omega t}\|=+\infty$, so that u does not belong to $C_\omega([0,+\infty[;\ X)$. By contradiction, (ii) holds.

(ii)$\Rightarrow$(i). Assume by contradiction that (i) does not hold. Due to Proposition 2.1, this means that the map

$$\gamma: C_\omega([0,+\infty[;\ Y)\to P_2(0)(X), \qquad \gamma f=\int_0^{+\infty}G(0,s)P_2(s)\Phi(s)f(s)\,ds,$$

is not onto. Since $P_2(0)(X)$ is finite-dimensional (by assumption (3.5)), γ is not onto if and only if

$$\gamma^*:(P_2(0)(X))^*\to(C_\omega([0,+\infty[;\ Y))^*$$

is not one to one, i.e., there is $x^*\in P_2(0)^*(X^*)\backslash\{0\}$ such that $\Phi(s)^*P_2(s)^*\times G(0,s)^*x^*=\Phi(s)^*G(0,s)^*x^*=0$ for $s\geqq 0$. Setting $s=\sigma+nT$, we get $\Phi(\sigma)^*G(0,\sigma)^*\times(V(0)^*)^{-n}x^*=0$, for $0\leqq\sigma\leqq T$ and $n\in\mathbf{N}$, and hence, for every $\xi\in\mathbf{C}$ with $|\xi|>\|(V(0)^*_{|P_2(0)^*(x^*)})^{-1}\|_{L(P_2(0)^*(X^*))}$ we have

$$\Phi(\sigma)^*G(0,\sigma)^*(\xi-(V(0)^*_{|P_2(0)^*(X^*)})^{-1})^{-1}x^*=0, \qquad 0\leqq\sigma\leqq T. \tag{3.8}$$

By assumption (3.5), the set $\{\xi\in\sigma(V(0)^*_{|P_2(0)^*(X^*)})^{-1};\ |\xi|>e^{-\omega T}\}$ is finite. Since $\xi\to(\xi-V(0)^*_{|P_2(0)^*(X^*)})^{-1}$ is holomorphic in its domain of definition, then (3.8) holds for every ξ in the resolvent set of $(V(0)^*_{|P_2(0)^*(X^*)})^{-1}$, with $|\xi|>e^{-\omega T}$. Then by (3.8) we get, for $\lambda\in\rho(V(0)^*)$, $|\lambda|<e^{-\omega T}$:

$$\Phi(\sigma)^*G(0,\sigma)^*(\lambda-V(0)^*)^{-1}(V(0)^*_{|P_2(0)^*(X^*)})^{-1}x^*=0, \qquad 0\leqq\sigma\leqq T. \tag{3.9}$$

Again using assumption (3.5), we find that (3.9) holds for every $\lambda\in\rho(V(0)^*)$. Let $\sigma_2=\{\lambda_1,\cdots,\lambda_N\}$. Then, since $x^*\in P_2(0)^*(X^*)$, we have $x^*=P_2(0)^*x=\sum_{i=1}^N P_{\lambda_i}(0)^*x^*$, where $P_{\lambda_i}(s)=1/2\pi i\int_{C(\lambda_i,\varepsilon)}(\lambda-V(s))^{-1}\,d\lambda$, $i=1,\cdots,N$, and $\varepsilon>0$ is sufficiently small. Since $x^*\neq 0$, there is $i\in\{1,\cdots,N\}$ such that $P_{\lambda_i}(0)^*x^*\neq 0$. Since λ_i is an eigenvalue with finite algebraic multiplicity, there is $m\in\mathbf{N}$ such that $(\lambda_i-V(0)^*)^mP^*_{\lambda_i}x^*=0$ and $y^*\overset{\text{def}}{=}(\lambda_i-V(0)^*)^{m-1}P^*_{\lambda_i}x^*\neq 0$. y^* belongs obviously to ker $(\lambda_i-V(0)^*)$, and from (3.9) we get, multiplying by λ^k $(k\in\mathbf{Z})$ and integrating over $C(\lambda_i,\varepsilon)$,

$$\Phi(\sigma)^*G(0,\sigma)^*(V(0)^*)^kP^*_{\lambda_i}x^*=0\quad\text{for each } k\in\mathbf{Z}$$

so that

$$\Phi(\sigma)^*G(0,\sigma)^*y^*=\Phi(\sigma)^*G(0,\sigma)^*(\lambda_i-V(0)^*)^{m-1}P^*_{\lambda_i}x^*=0.$$

This contradicts assumption (ii). Therefore (ii) implies (i).

Now let one of the equivalent conditions (i) or (ii) hold. Arguing as in the proof of (ii)⇒(i), we can see that the mapping

$$\Gamma: C^\alpha_\omega([0,+\infty[;\,Y)\to P_2(0)(X),\quad \Gamma f=\int_0^{+\infty}G(0,s)P_2(s)\Phi(s)f(s)\,ds,$$

is onto. The statement now follows applying Proposition 2.1. □

3.2. Applications to distributed control problems. Here we consider the problem

$$\begin{aligned} &u_t(t,x)=\mathscr{A}(t,x,\partial)u(t,x)+y_k(t)\phi_k(t,x), && t>0,\quad x\in\Omega,\\ &u(0,x)=u_0(x), && x\in\Omega, \\ &\mathscr{B}u(t,x)=0, && \end{aligned}\tag{3.10}$$

where the linear operator $\mathscr{A}(t,x,\partial)$ is given by

$$\mathscr{A}(t,\cdot,\partial)f=a_{ij}(t,\cdot)D_{ij}f+b_i(t,\cdot)D_if+c(t,\cdot)f,\tag{3.11}$$

and

$$\begin{aligned} &\text{(a)}\quad a_{ij}(t,x)=a_{ji}(t,x),\qquad a_{ij}(t,x)\xi_i\xi_j\geqq\nu|\xi|^2,\\ &\text{(b)}\quad a_{ij}(t+T,x)=a_{ij}(t,x),\quad b_i(t+T,x)=b_i(t,x),\quad c(t+T,x)=c(t,x),\\ &\text{(c)}\quad t\to a_{ij}(t,\cdot),\quad t\to b_i(t,\cdot),\quad t\to c(t,\cdot)\in C^\alpha(\mathbf{R};\,C(\bar\Omega)). \end{aligned}\tag{3.12}$$

The boundary operator $\mathscr{B}$ is either the trace operator on $\partial\Omega$ or else any mixed nontangential differential operator:

$$(\mathscr{B}g)(x)=\beta_i(x)D_ig(x)+\gamma(x)g(x),\qquad x\in\partial\Omega,\tag{3.13}$$

with

$$\beta_i,\ \gamma\in C^1(\partial\Omega),\quad \beta_i(x)\nu_i(x)\neq 0,\quad x\in\partial\Omega.\tag{3.14}$$

We choose $p>1$ and we set

$$X=L^p(\Omega),\qquad D=\{\phi\in W^{2,p}(\Omega);\ \mathscr{B}\phi=0\}.\tag{3.15}$$

Then the operators

$$A(t): D\to X,\qquad A(t)\phi=\mathscr{A}(t,x,\partial)\phi\tag{3.16}$$

satisfy assumption (2.2) due to [3] and [2]. The evolution operator $G(t,s)$ generated by $\{A(t);\,t\in\mathbf{R}\}$ is compact because D is compactly embedded in X; therefore the spectrum of $V(s)$ consists of isolated eigenvalues (except the point $z=0$) with finite algebraic multiplicity, and it is independent of s. Hence, assumptions (2.10), (3.5) are satisfied by any $\omega \geqq 0$ such that no complex number with modulus $e^{-\omega T}$ is an eigenvalue of $V(0)$.

We choose

$$Y=\mathbf{R}^k, \qquad \Phi(t)(y_1,\cdots,y_k)=\phi_i(t,\cdot)y_i, \tag{3.17}$$

and we assume that $\phi_1,\cdots,\phi_k$ are linearly independent and such that

$$t\to\phi_i(t,\cdot)\in C^\alpha(\mathbf{R};X), \quad \phi_i(t+T,\cdot)=\phi_i(t,\cdot) \quad \forall t\in\mathbf{R}. \tag{3.18}$$

For every $\lambda\in\sigma(V(0)^*)$ let $\xi_1^*,\cdots,\xi_{N(\lambda)}^*$ be linearly independent functions in $L^{p^*}(\Omega)$ $(1/p+1/p^*=1)$ spanning the eigenspace of $V(0)^*$ with eigenvalue λ. Then for every $s\in[0,T]$ the mapping $\Phi(s)^*G(T,s)^*:\ker(\lambda-V(0)^*)\to(\mathbf{R}^k)^*=\mathbf{R}^k$ may be represented (with respect to such a basis) by the matrix

$$[A_{ji}(s)]=[\langle G(T,s)\phi_j(s,\cdot),\xi_i^*\rangle]_{i=1,\cdots,N(\lambda),j=1,\cdots,k}. \tag{3.19}$$

Hence, the result of Theorem 3.1 can be reformulated as follows.

PROPOSITION 3.2. *Fix any $\omega\geqq 0$ such that no eigenvalue of $V(0)$ has modulus $e^{-\omega T}$. Then the following statements are equivalent:*

(i) *For each $u_0\in L^p(\Omega)$ there are $y_1,\cdots,y_k\in C_\omega^\alpha([0,+\infty[;\mathbf{R})$ such that the solution u of* (3.10) *satisfies*

$$\sup_{t\geqq 0}\|e^{\omega t}u(t,\cdot)\|_{L^p(\Omega)}<+\infty, \quad \sup_{t\geqq a}\|e^{\omega t}u(t,\cdot)\|_{W^{2,p}(\Omega)}<+\infty \quad \forall a>0;$$

(ii) *For each $\lambda\in\sigma(V(0))$ with $|\lambda|>e^{-\omega T}$, and for each $y\in\mathbf{R}^{N(\lambda)}\backslash\{0\}$ there is $s\in[0,T]$ such that $[A_{ji}(s)]y\neq 0$, where $[A_{ji}(s)]$ is the matrix defined in* (3.19). □

From Proposition 3.2 we get the following sufficient condition for the stabilizability of system (3.10).

COROLLARY 3.3. *The following conditions are equivalent:*

(i) *For each $\lambda\in\sigma(V(0))$ with $|\lambda|>e^{-\omega T}$, we have $N(\lambda)\leqq k$, and there is $\bar{s}\in[0,T]$ such that the rank of the matrix $[A_{ji}(\bar{s})]$ defined in* (3.19) *is $N(\lambda)$;*

(ii) *For each $\lambda\in\sigma(V(0))$ with $|\lambda|>e^{-\omega T}$ we have $N(\lambda)\leqq k$, and there is $\bar{s}\in[0,T]$ such that, denoting by $\{\chi_1^*,\cdots,\chi_{N(\lambda)}^*\}$ any system of generators of* $\ker(\lambda-V(\bar{s})^*)$, *the rank of the matrix*

$$[B_{ji}]=[\langle\phi_j(\bar{s},\cdot),\chi_i^*\rangle]_{i=1,\cdots,N(\lambda),j=1,\cdots,k}$$

is $N(\lambda)$.

If either (i) *or* (ii) *holds, then conditions* (i) *and* (ii) *of Proposition* 3.2 *hold, so that system* (3.10) *is stabilizable.*

Proof. Since $G(T,s)^*:\ker(\lambda-V(0)^*)\to\ker(\lambda-V(s)^*)$ is an isomorphism, then $\{\xi_1^*,\cdots,\xi_{N(\lambda)}^*\}$ is a basis of $\ker(\lambda-V(0)^*)$ if and only if $\{G(T,s)^*\xi_1^*,\cdots,G(T,s)^*\xi_{N(\lambda)}^*\}$ is a basis of $\ker(\lambda-V(s)^*)$. Therefore, conditions (i) and (ii) are clearly equivalent. Moreover, if (i) holds, then for every $y\in\mathbf{R}^{N(\lambda)}\backslash\{0\}$ we have $[A_{ji}(\bar{s})]y\neq 0$, so that (ii) of Proposition 3.2 holds. □

4. Stabilizability in boundary control problems. Here we consider a parabolic initial value problem nonhomogeneous at the boundary,

$$\begin{aligned} &v_t(t,x)=\mathscr{A}(t,x,\partial)v(t,x), && t>0, \quad x\in\Omega,\\ &v(0,x)=v_0(x), && x\in\Omega,\\ &\mathscr{B}(x,\partial)v(t,x)=\Psi(t)g(t)(x), && t>0, \quad x\in\partial\Omega, \end{aligned} \tag{4.1}$$

where, as before, Ω is a bounded open set in $\mathbf{R}^n$ with C^2 boundary $\partial\Omega$, the linear operator $\mathscr{A}(t, x, \partial)$ is defined in (3.11) and its coefficients satisfy assumption (3.12), and $\mathscr{B}(x, \partial)$ is either the trace operator on $\partial\Omega$ or the differential operator defined in (3.13), with coefficients satisfying (3.14), $g(\cdot)$ is a function with values in a Banach space Z, and $\Psi(t)$ is a linear (time periodic) operator in $L(Z, L^p(\partial\Omega))$, for some $p > 1$.

$$(4.2)\quad \begin{aligned} &\text{(a)}\quad t \to a_{ij}(t, \cdot),\ t \to b_i(t, \cdot), \qquad t \to c(t, \cdot) \in C^{1+\alpha}(\mathbf{R};\ C(\bar{\Omega})),\\ &\text{(b)}\quad x \to a_{ij}(t, x) \in C^2(\bar{\Omega}),\quad x \to b_i(t, x) \in C^1(\bar{\Omega}),\quad x \to c(t, x) \in C(\bar{\Omega}),\quad \forall t \in \mathbf{R},\\ &\text{(c)}\quad \sup_{t\in\mathbf{R}} \|a_{ij}(t, \cdot)\|_{C^2(\bar{\Omega})} + \sup_{t\in\mathbf{R}} \|b_i(t, \cdot)\|_{C^1(\bar{\Omega})} + \sup_{t\in\mathbf{R}, x\in\bar{\Omega}} |c(t, x)| < +\infty, \end{aligned}$$

and we choose, as before, $X = L^p(\Omega)$, $D = \{\phi \in W^{2,p}(\Omega);\ \mathscr{B}\phi = 0\}$. Due to assumption (4.2)(a), the function $t \to A(t)$ (defined in (3.16)) belongs to $C^{1+\alpha}(\mathbf{R};\ L(D, X))$, so that there exists $G_s(t, s) \in L(X, D)$ for $t > s$ (see [17]).

4.1. The Dirichlet boundary condition. In this case, D is the space $W^{2,p}(\Omega) \cap W_0^{1,p}(\Omega)$. We assume that

$$(4.3)\qquad 0 \in \rho(A(t)) \quad \forall t \in \mathbf{R},$$

so that for each $t \in \mathbf{R}$ the Dirichlet mapping $D(t) \in L(W^{2-1/p,p}(\partial\Omega), W^{2,p}(\Omega))$ is well defined by $D(t)y = z$, where z is the unique solution of the elliptic boundary value problem

$$(4.4)\qquad \mathscr{A}(t, x, \partial)z = 0 \quad \text{in } \Omega, \qquad z|_{\partial\Omega} = y$$

(see [3]). Concerning the linear operators $\Psi(t)$, we assume that

$$(4.5)\qquad \begin{aligned} &t \to \Psi(t) \in C^1(\mathbf{R}, L(Z, L^p(\partial\Omega))) \cap C(\mathbf{R}, L(Z, W^{2-1/p,p}(\partial\Omega))),\\ &\hspace{16em}\Psi(t + T) = \Psi(t), \quad t \in \mathbf{R}. \end{aligned}$$

In [17] we showed that for every $\xi \in C^1([0, +\infty[;\ L^p(\partial\Omega)) \cap C([0, +\infty[;\ W^{2-1/p,p}(\partial\Omega))$ the function $D(\cdot)\xi(\cdot)$ belongs to $C^1([0, +\infty[;\ X_\theta)$ for each $\theta < 1/2p$, and, if v_0 belongs to $L^p(\Omega)$, then problem

$$w'(t) = \mathscr{A}(t, \cdot, \partial)w, \quad t > 0, \quad w(0, \cdot) = w_0, \quad \mathscr{B}w(t) = \xi(t), \quad t > 0$$

has a unique solution $w \in C([0, +\infty[;\ L^p(\Omega)) \cap C^1(]0, +\infty[;\ L^p(\Omega)) \cap C(]0, +\infty[;\ W^{2,p}(\Omega))$, given by the representation formula

$$\begin{aligned} W(t) &= G(t, 0)w_0 + \int_0^t G_s(t, s)D(s)\xi(s)\, ds\\ &= G(t, 0)(w_0 - D(0)\xi(0)) - \int_0^t G(t, s)\frac{\partial}{\partial s}[D(s)\xi(s)]\, ds + D(t)\xi(t), \qquad t \geqq 0. \end{aligned}$$

Therefore, if g belongs to $C^1([0, +\infty[;\ Z)$, problem (4.1) has a unique solution v such that $t \to v(t, \cdot)$ belongs to $C([0, +\infty[;\ L^p(\Omega)) \cap C^1(]0, +\infty[;\ L^p(\Omega)) \cap C(]0, +\infty[;\ W^{2,p}(\Omega))$, and v is given by

$$\begin{aligned} v(t, \cdot) &= G(t, 0)v_0 + \int_0^t G_s(t, s)D(s)\Psi(s)g(s, \cdot)\, ds\\ (4.6)\qquad &= G(t, 0)(v_0 - D(0)\Psi(0)g(0, \cdot)) - \int_0^t G(t, s)\frac{\partial}{\partial s}[D(s)\psi(s)g(s, \cdot)]\, ds\\ &\quad + D(t)\Psi(t)g(t, \cdot), \qquad t \geqq 0. \end{aligned}$$

We want to find sufficient conditions on the data to ensure that for every $v_0 \in X$ there is $g \in C^1([0, +\infty[; Z)$ such that $v(t, \cdot)$ decays exponentially as $t \to +\infty$ in the $W^{2,p}$ norm. We use notation from § 2.

THEOREM 4.1. *Let* (4.2), $\cdots$, (4.5), (3.12)(*a*) *hold, and let* $\omega \geqq 0$ *be such that* $\sigma(V(s)) \cap \{\lambda \in \mathbf{C}; |\lambda| = e^{-\omega T}\} = \varnothing$ *for each* $s \in \mathbf{R}$. *Then the following statements are equivalent*:

(i) *For every* $v_0 \in L^p(\Omega)$ *there is* $g \in C^1_\omega([0, +\infty[; Z)$ *such that* $t \to v(t, \cdot)$ (*where* v *is the solution of* (4.1)) *belongs to* $C_\omega([0, +\infty[; L^p(\Omega))$;

(ii) *For every* $\lambda \in \mathbf{C}$ *with* $|\lambda| > e^{-\omega T}$ *we have*

$$(\lambda - V(0)^*)x^* = 0, \quad \Psi(s)^*\left[\frac{\partial}{\partial \nu_s}(G(T, s)^*x^*)\right] = 0 \quad \textit{for each } s \in [0, T] \Rightarrow x^* = 0,$$

where $\partial/\partial \nu_s \psi = a_{ij}(s, \cdot)D_i\psi(\cdot)\nu_j(\cdot)$ *and we identify* $G(T, s)^*x^*$ *with the function* $f \in W^{2,p^*}(\Omega)$ *such that* $(G(T, s)^*x^*)(y) = \int_\Omega y(x)f(x)\,dx$ *for each* $y \in L^p(\Omega)$.

If one of the equivalent conditions (i) *or* (ii) *holds, then for every* $v_0 \in L^p(\Omega)$ *there is* $g \in C^1_\omega([0, +\infty[; Z)$ *such that* $t \to v(t, \cdot)$ *belongs to* $C_\omega([a, +\infty[; W^{2,p}(\Omega))$ *for each* $a > 0$.

Proof. The proof is quite analogous to that of Theorem 3.1, therefore some steps will be only sketched.

First, from [17] it follows that if $t > s$, then $G(t, s)^*x^*$ belongs to $D(A(s)^*) \overset{\text{def}}{=} \{y^* \in X^*$: the mapping $D \to C$, $x \to \langle A(s)x, y^*\rangle$ has a continuous extension to $X\}$, and

$$G_s(t, s)^*x^* = -A(s)^*G(t, s)^*x^*.$$

Moreover, a simple integration by parts argument shows that for each $y^* \in D(A(s)^*) = W^{2,p^*}(\Omega) \cap W_0^{1,p^*}(\Omega)$ we have $A(s)^*y^* = D_{ij}(a_{ij}(s, \cdot)y^*(\cdot)) - D_j(b_j(\cdot)y^*(\cdot)) + c(\cdot)y^*(\cdot)$, and

$$D(s)^*A(s)^*y^* = \frac{\partial}{\partial \nu_s} y^*.$$

Therefore

$$D(s)^*(G_s(t, s)^*x^* = -\frac{\partial}{\partial \nu_s}(G(t, s)^*x^*), \qquad t > s. \tag{4.7}$$

Due to equality (4.7) and the first representation formula in (4.6), the proof of (i)$\Leftrightarrow$(ii) is the same (with obvious modifications) as the corresponding one in Theorem 3.1.

Now let one of the equivalent conditions (i) or (ii) hold; then for every $u_0 \in L^p(\Omega)$ there is $g \in C^1_\omega([0, +\infty[; Z)$ such that $\limsup_{t \to +\infty} \|e^{\omega t}u(t, \cdot)\|_{L^p(\Omega)} < +\infty$. We show that also $\limsup_{t \to +\infty} \|e^{\omega t}u(t, \cdot)\|_{W^{2,p}(\Omega)} < +\infty$.

Since $t \to v(t, \cdot)$ belongs to $C_\omega([0, +\infty[; L^p(\Omega))$, then by Proposition 2.3 we have

$$P_2(0)v_0 = -\int_0^{+\infty} \frac{\partial}{\partial s}[G(0, s)P_2(s)]D(s)\Psi(s)g(s)\,ds.$$

As remarked before, $s \to D(s)\Psi(s)g(s)$ belongs to $C^1([0, +\infty[; X_\theta)$ for each $\theta < 1/2p$; therefore, by Corollary 2.6 of [17], we can integrate by parts to get

$$P_2(0)[v_0 - D(0)\Psi(0)g(0)] = \int_0^{+\infty} G(0, s)P_2(s)\frac{\partial}{\partial s}[D(s)\Psi(s)g(s)]\,ds. \tag{4.8}$$

Since $s \to \Psi(s)g(s)$ belongs to $C^1([0, +\infty[; L^p(\partial\Omega)) \cap C([0, +\infty[; W^{2-1/p,p}(\partial\Omega))$, then, by using Proposition 3.1 of [17], we can see easily that $v \to v(t, \cdot)$ belongs to

$C([0,+\infty[;\ L^p(\Omega))\cap C^1(]0,+\infty[;\ L^p(\Omega))\cap C(]0,+\infty[;\ W^{2,p}(\Omega))$, and it may be represented by the second formula in (4.6). Since $s\to D(t)\Psi(t)g(t)$ belongs to $C_\omega([0,+\infty[;\ W^{2,p}(\Omega))$, then $t\to v(t,\cdot)$ is in $C_\omega([a,+\infty[;\ W^{2,p}(\Omega))$ for every $a>0$ if and only if the function

$$z(t)=G(t,0)[v_0-D(0)\Psi(0)g(0)]-\int_0^t G(t,s)\frac{\partial}{\partial s}[D(s)\Psi(s)g(s)]\,ds,\qquad t\geqq 0,$$

does. Since $d/ds\, D(s)\Psi(s)g(s)$ belongs to $C_\omega([0,+\infty[;\ X_\theta)$ with $\theta>0$ and (4.8) holds, the last statement of Proposition 2.1 implies $z\in C_\omega([a,+\infty[;\ W^{2,p}(\Omega))$ for every $a>0$. □

THEOREM 4.2. *Under the previous assumptions and notation, let* $Z=W^{2-1/p,p}(\partial\Omega)$, $\Psi(s)=\mathbf{I}$ *for each* s. *Then system* (4.1) *is stabilizable, provided* $\mathscr{A}(t,\cdot,\partial)$ *is formally self-adjoint, and* $\partial\Omega$ *and the coefficients of* $\mathscr{A}(t,\cdot,\partial)$ *are of class* C^∞ *for every* t.

Proof. For each $x^*\in X^*$, $G(T,s)^*x^*$ is the solution of

$$u'(s)=-A(s)^*u(s),\quad s<T,\qquad u(T)=x^*,$$

so that, setting $t=T-s$, we have $G(T,s)^*x^*=v(T-s)$, where v satisfies

$$v'(t)=A(T-t)^*v(t),\quad t>0,\qquad v(0)=x^*.$$

Now, a result of [19] ensures that the set $\{(t,x)\in[0,T]\times\partial\Omega;\ v(t,x)=\partial/\partial\nu\, v(t,x)=0\}$ has Lebesgue measure zero. Therefore $s\to\partial/\partial\nu_s(G(T,s)^*x^*)$ cannot vanish identically in $[0,T]$, and we can apply Theorem 4.1, which implies the statement. □

4.2. Applications to boundary control problems (I). Under the previous assumptions and notation, we consider in particular the case where

$$Z=\mathbf{R}^k,\qquad \Psi(t)(y_1,\cdots,y_k)=\psi_i(t,\cdot)y_i,\tag{4.9}$$

and

$$t\to\psi_i(t,\cdot)\in C^1(\mathbf{R};\ L^p(\partial\Omega))\cap C(\mathbf{R};\ W^{2-1/p,p}(\partial\Omega)),\qquad \psi_i(t+T,\cdot)=\psi_i(t,\cdot)\quad\forall t\in\mathbf{R}.\tag{4.10}$$

Applying Theorem 4.1 we get results similar to the ones of Proposition 3.2 and Corollary 3.3; the proof is also similar and it is omitted.

PROPOSITION 4.3. *Let* (4.2), (4.3), (4.9), (4.10) *hold, and let* $\omega\geqq 0$ *be such that no eigenvalue of* $V(0)$ *has modulus* $e^{-\omega T}$. *Then the following statements are equivalent*:

(i) *For each* $v_0\in L^p(\Omega)$ *there are* $y_1,\cdots,y_k\in C^1_\omega([0,+\infty[;\ \mathbf{R})$ *such that the solution* v *of* (4.1) *satisfies*

$$\sup_{t\geqq 0}\|e^{\omega t}v(t,\cdot)\|_{L^p(\Omega)}<+\infty,\quad \sup_{t\geqq a}\|e^{\omega t}v(t,\cdot)\|_{W^{2,p}(\Omega)}<+\infty\quad\forall a>0;$$

(ii) *For each* $\lambda\in\sigma(V(0))$ *with* $|\lambda|>e^{-\omega T}$, *denote by* $\{\xi_1^*,\cdots,\xi_{N(\lambda)}^*\}$ *any basis of* $\ker(\lambda-V(0)^*)$ *and set*

$$[A_{ji}(s)]=\left[\left\langle\psi_j(s,\cdot),\frac{\partial}{\partial\nu_s}G(T,s)^*\xi_i^*\right\rangle\right]_{i=1,\cdots,N(\lambda),j=1,\cdots,k}.\tag{4.11}$$

Then for every $y\in\mathbf{R}^{N(\lambda)}\backslash\{0\}$, *there is* $s\in[0,T]$ *such that* $[A_{ji}(s)]y\neq 0$.

Condition (i) *and* (ii) *hold if one of the following equivalent conditions is satisfied*:

(iii) *For each* $\lambda\in\sigma(V(0))$ *with* $|\lambda|>e^{-\omega T}$ *we have* $N(\lambda)\leqq k$, *and there is* $\bar{s}\in[0,T]$ *such that the rank of the matrix* $[A_{ji}(\bar{s})]$ *is* $N(\lambda)$;

(iv) *For each* $\lambda \in \sigma(V(0))$ *with* $|\lambda| > e^{-\omega T}$ *we have* $N(\lambda) \leqq k$, *and there is* $\bar{s} \in [0, T]$ *such that, denoting by* $\{\chi_1^*, \cdots, \chi_{N(\lambda)}^*\}$ *any system of generators of* $\ker(\lambda - V(\bar{s})^*)$, *the rank of the matrix*

$$[B_{ji}] = \left[\left\langle \psi_j(\bar{s}, \cdot), \frac{\partial}{\partial \nu_s} \chi_i^* \right\rangle\right]_{i=1,\cdots,N(\lambda), j=1,\cdots,k}$$

is $N(\lambda)$. □

4.3. The mixed boundary condition. The space D is now $D = \{g \in W^{2,p}(\Omega); \mathscr{B}(x, \partial)g = 0\}$, where $\mathscr{B}$ is the differential operator defined in (3.13). Assuming again that (4.3) holds, we may define the mixed mapping $M(t)$ by $M(t)g = z$, where z is the unique solution of the elliptic boundary value problem (1.6).

The assumptions on the linear operators $\Psi(t)$ are the following:

$$\text{(4.12)} \qquad \begin{aligned} &t \to \Psi(t) \in C^\beta(\mathbf{R}, L(Z, L^p(\partial\Omega))) \cap C(\mathbf{R}, L(Z, W^{1-1/p,p}(\partial\Omega))), \\ &\Psi(t+T) = \Psi(t), \quad t \in \mathbf{R}, \end{aligned}$$

where

$$\text{(4.13)} \qquad \beta > \frac{1}{2} + \frac{1}{2p}.$$

The linear mapping $M(t)$ belongs to $L(L^p(\partial\Omega), W^{1-1/p-\varepsilon,p}(\Omega)) \cap L(W^{1-1/p,p}(\partial\Omega), W^{2,p}(\Omega))$ for each $t \in \mathbf{R}$ and $\varepsilon > 0$ (see [3] and [14]; in fact, in [14] more regularity assumptions are made in order to treat a larger class of boundary value problems, but the reader can check that in this regularity problem our hypotheses are sufficient). Moreover, assumption (4.2) implies also that $t \to M(t)$ belongs to $C^1(\mathbf{R}; L(L^p(\partial\Omega), W^{1-1/p-\varepsilon,p}(\Omega))) \cap C^1(\mathbf{R}; L(W^{1-1/p,p}(\partial\Omega), W^{2,p}(\Omega)))$. On the other hand, by [10] we have

$$X_\theta = B_\infty^{2\theta,p}(\Omega) \quad \text{for } \theta < \frac{1}{2} - \frac{1}{2p},$$

with equivalence of the respective norms, so that, recalling that $W^{1-1/p-\varepsilon,p}(\Omega)$ is continuously embedded in $B_\infty^{1-1/p-\varepsilon,p}(\Omega)$, we also get

$$t \to M(t) \in C^1(\mathbf{R}; L(Z, X_\theta)).$$

Therefore, for every $\psi \in C^\beta([0, +\infty[; L^p(\partial(\Omega)) \cap C([0, +\infty[; W^{1-1/p,p}(\partial\Omega))$, we have

$$t \to M(t)\psi(t) \in C^\beta([0, +\infty[; X_\theta) \cap C([0, +\infty[; W^{2,p}(\Omega)), \qquad \theta < \frac{1}{2} - \frac{1}{2p}.$$

Under such regularity assumptions, it can be deduced from [17] that, for every $w_0 \in L^p(\Omega)$, the problem

$$w'(t) = \mathscr{A}(t, \cdot, \partial)w, \quad t > 0, \quad w(0, \cdot) = w_0, \quad \mathscr{B}w(t) = \psi(t), \quad t > 0,$$

has a unique solution $w \in C([0, +\infty[; L^p(\Omega)) \cap C^1(]0, +\infty[; L^p(\Omega)) \cap C([0, +\infty[; W^{2,p}(\Omega))$, given by the representation formula

$$w(t) = G(t, 0)w_0 + \int_0^t G_s(t, s)M(s)\psi(s)\, ds, \qquad t \geqq 0.$$

Therefore, if $g:[0,+\infty[\to Z$ is locally β-Hölder continuous, then problem (4.1) has a unique solution v such that $t\to v(t,\cdot)\in C([0,+\infty[;\ L^p(\Omega))\cap\ C^1(]0,+\infty[;\ L^p(\Omega))\cap C([0,+\infty[;\ W^{2,p}(\Omega))$, given by

$$v(t,\cdot)=G(t,0)v_0+\int_0^t G_s(t,s)M(s)\Psi(s)g(s,\cdot)\,ds,\qquad t\geqq 0.$$

The following result, concerning stabilizability for problem (4.1), is quite similar to the one of Theorem 4.1.

THEOREM 4.4. *Let* (4.2), (4.3), (4.12), (4.13), (3.12)(a) *hold, and let* $\omega\geqq 0$ *be such that* $\sigma(V(s))\cap\{\lambda\in\mathbf{C};\ |\lambda|=e^{-\omega T}\}=\emptyset$ *for each* $s\in\mathbf{R}$. *Then the following statements are equivalent*:

(i) *For every* $v_0\in L^p(\Omega)$ *there is* $g\in C^\beta_\omega([0,+\infty[;\ Z)$ *such that* $t\to v(t,\cdot)$ (*where* v *is the solution of* (4.1)) *belongs to* $C_\omega([0,+\infty[;\ L^p(\Omega))$;

(ii) *For every* $\lambda\in\mathbf{C}$ *with* $|\lambda|>e^{-\omega T}$ *we have*

$$(\lambda-V(0)^*)x^*=0,\quad \Psi(s)^*[(G(T,s)^*x^*)_{|\partial\Omega}]=0\quad \textit{for each } s\in[0,T]\Rightarrow x^*=0,$$

where we identify $G(T,s)^*x^*$ *with the function* $f\in W^{2,p^*}(\Omega)$ *such that* $(G(T,s)^*x^*)(y)=\int_\Omega y(x)f(x)\,dx$ *for each* $y\in L^p(\Omega)$.

If one of the equivalent conditions (i) *or* (ii) *holds, then for every* $v_0\in L^p(\Omega)$ *there is* $g\in C^\beta_\omega([0,+\infty[;Z)$ *such that* $t\to v(t,\cdot)$ *belongs to* $C_\omega([a,+\infty[;\ W^{2,p}(\Omega))$ *for each* $a>0$.

Proof. It is similar to the one of Theorem 4.1. Concerning $D(A(s)^*)$, it is well known (see [14]) that there are a first-order boundary operator $\mathscr{B}^*=\mathscr{B}^*(s,\cdot,\partial)$, and a zero order boundary operator $\mathscr{C}^*$ (with $\mathscr{C}^*\phi=c^*(s,x)\phi(x)$, $c^*(s,x)\neq 0$ for any $x\in\partial\Omega$) such that for every couple of regular functions u, v we have

$$\int_\Omega [v(x)(\mathscr{A}(s,x,\partial)u)(x)-u(x)(\mathscr{A}^*(s,x,\partial)v)(x)]\,dx$$

$$=\int_{\partial\Omega}[u(\cdot)\mathscr{B}^*v(\cdot)-\mathscr{B}u(\cdot)\mathscr{C}^*v(\cdot)]\,d\sigma_x,$$

where $\mathscr{A}^*(s,\cdot,\partial)$ is defined in (4.7). Then $D(A(s)^*)=\{y^*\in W^{2,p^*}(\Omega):\ \mathscr{B}^*y^*=0\}$, and $A(s^*)y^*=\mathscr{A}^*(s,\cdot,\partial)y^*$. Integrating by parts we easily get

$$(M(s)^*A(s)^*y^*)(x)=-c^*(s,x)y^*(x),\quad s\in\mathbf{R},\quad x\in\partial\Omega,\quad y^*\in D(A(s)^*),$$

so that

$$M(s)^*G_s(t,s)^*x^*=c^*(s,\cdot)G(t,s)^*x^*_{|\partial\Omega},\quad \forall x^*\in X^*=L^{p^*}(\Omega),\quad t>s,$$

and we have to apply Proposition 2.3 instead of Proposition 2.1. □

4.4. Applications to boundary control problems (II). We again consider the case where

$$Z=\mathbf{R}^k,\qquad \Psi(t)(y_1,\cdots,y_k)=\psi_i(t,\cdot)y_i,\tag{4.14}$$

and, for some $\beta\in]\frac{1}{2},1[$

$$\begin{aligned}&t\to\psi_i(t,\cdot)\in C^\beta(\mathbf{R};\ L^p(\partial\Omega))\cap C(\mathbf{R};\ W^{2-1/p,p}(\partial\Omega)),\\ &\qquad\qquad\qquad\qquad\qquad\phi_i(t+T,\cdot)=\phi_i(t,\cdot)\quad\forall t\in\mathbf{R}.\end{aligned}\tag{4.15}$$

Applying Theorem 4.1 we get a result similar to the one of Proposition 3.2.

PROPOSITION 4.5. *Let* (3.12)(a), (4.2), (4.3), (4.14), (4.15) *hold, and let* $\omega\geqq 0$ *be such that no eigenvalue of* $V(0)$ *has modulus* $e^{-\omega T}$. *Then the following statements are equivalent*:

(i) *For each* $v_0 \in L^p(\Omega)$ *there are* $y_1, \cdots, y_k \in C^\beta_\omega([0, +\infty[; \mathbf{R})$ *such that the solution* v *of* (4.1) *satisfies*

$$\sup_{t \geqq 0} \|e^{\omega t} v(t, \cdot)\|_{L^p(\Omega)} < +\infty, \quad \sup_{t \geqq a} \|e^{\omega t} v(t, \cdot)\|_{W^{2,p}(\Omega)} < +\infty \quad \forall a > 0;$$

(ii) *For each* $\lambda \in \sigma(V(0))$ *with* $|\lambda| > e^{-\omega T}$, *denote by* $\{\xi_1^*, \cdots, \xi_{N(\lambda)}^*\}$ *any basis of* $\ker(\lambda - V(0)^*)$ *and set*

$$[A_{ji}(s)] = [\langle \psi_j(s, \cdot), (G(T, s)^* \xi_i^*)_{|\partial\Omega} \rangle]_{i=1,\cdots,N(\lambda), j=1,\cdots,k}. \tag{4.16}$$

Then for every $y \in \mathbf{R}^{N(\lambda)} \backslash \{0\}$, *there is* $s \in [0, T]$ *such that* $[A_{ji}(s)] y \neq 0$.

Conditions (i) *and* (ii) *hold if one of the following equivalent conditions is satisfied*:

(iii) *For each* $\lambda \in \sigma(V(0))$ *with* $|\lambda| > e^{-\omega T}$ *we have* $N(\lambda) \leqq k$, *and there is* $\bar{s} \in [0, T]$ *such that the rank of the matrix* $[A_{ji}(\bar{s})]$ *defined in* (4.16) *is* $N(\lambda)$;

(iv) *For each* $\lambda \in \sigma(V(0))$ *with* $|\lambda| > e^{-\omega T}$ *we have* $N(\lambda) \leqq k$, *and there is* $\bar{s} \in [0, T]$ *such that, denoting by* $\{\chi_1^*, \cdots, \chi_{N(\lambda)}^*\}$ *any system of generators of* $\ker(\lambda - V(\bar{s})^*$, *the rank of the matrix*

$$[B_{ji}] = [\langle \psi_j, \chi_{i|\partial\Omega} \rangle]_{i=1,\cdots,N(\lambda), j=1,\cdots,k}$$

is $N(\lambda)$. □

Acknowledgments. The author wishes to thank the referees for their careful reading of the manuscript and their useful suggestions.

REFERENCES

[1] P. Acquistapace, F. Flandoli, and B. Terreni, *Boundary control of nonautonomous parabolic systems*, SIAM J. Control Optim., to appear.

[2] S. Agmon, *On the eigenfunctions and the eigenvalues of general elliptic boundary value problems*, Comm. Pure Appl. Math., 15 (1962), pp. 119-147.

[3] S. Agmon, A. Douglis, and L. Nirenberg, *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, Comm. Pure Appl. Math., 12 (1959), pp. 623-727.

[4] H. Amann, *Feedback Stabilization of Linear and Semilinear Parabolic Systems*, in Proc. Trends in Semigroup Theory and Applications, Trieste, Sept. 28-Oct. 2, 1987, Lecture Notes in Pure and Applied Mathematics, Marcel Dekker, New York, 1989, pp. 21-57.

[5] V. A. Balakrishnan, *Applied Functional Analysis*, 2nd ed., Springer-Verlag, New York, Heidelberg, Berlin, 1981.

[6] R. Datko, *Some nonautonomous control problems with quadratic cost*, J. Differential Equations, 21 (1976), pp. 231-262.

[7] G. Da Prato and A. Ichikawa, *Quadratic control for linear time varying systems*, SIAM J. Control Optim., to appear.

[8] G. Da Prato and J. P. Zolesio, *A boundary control problem for a parabolic equation in noncylindrical domain*, in Proc. COMCON Conference Inc., Workshop on stabilization of Flexible Structures, Jan. 23-26, 1988, Montpellier; Optimization Software Inc., New York, Los Angeles, 1987, pp. 52-61.

[9] F. Flandoli, *Riccati equation arising in a boundary control problem with distributed parameters*, SIAM J. Control Optim., 22 (1984), pp. 76-86.

[10] P. Grisvard, *Equations différentielles abstraites*, Ann. Sci. Ecole Norm. Sup. (4), 2 (1969), pp. 311-395.

[11] G. A. Hewer, *Periodicity, detectability and the matrix Riccati equation*, SIAM J. Control Optim., 13 (1975), pp. 1235-1251.

[12] I. Lasiecka and R. Triggiani, *Stabilization of Neumann boundary feedback of parabolic equations: the case of trace in the feedback loop*, Appl. Math. Optim., 10 (1983), pp. 307-350.

[13] ———, *Stabilization and structural assignment of Dirichlet boundary feedback parabolic equations*, SIAM J. Control Optim., 21 (1983), pp. 766-803.

[14] J. L. Lions and E. Magenes, *Problemi ai limiti non omogenei (V)*, Ann. Sci. Norm. Sup. Pisa, 16 (1962), pp. 1-44.

[15] A. LUNARDI, *On the evolution operator for abstract parabolic equations*, Israel J. Math., 60 (1987), pp. 281-314.

[16] ———, *Bounded solutions of linear periodic abstract parabolic equations*, Proc. Roy. Soc. Edinburgh, Sect. A, 110 (1988), pp. 135-159.

[17] ———, *Differentiability with respect to (t, s) of the parabolic evolution operator*, Israel J. Math., 68 (1989), pp. 161-184.

[18] T. NAMBU, *On the stabilization of diffusion equations*, J. Differential Equations, 52 (1984) pp. 204-233.

[19] G. SCHMIDT AND N. WECK, *On the boundary behavior of solutions to elliptic and parabolic equations—with applications to boundary control for parabolic equations*, SIAM J. Control Optim., 16 (1978), pp. 593-598.

[20] P. E. SOBOLEVSKII, *Equations of parabolic type in a Banach space*, Amer. Math. Soc. Transl., 49 (1966), pp. 1-62.

[21] R. TRIGGIANI, *Boundary feedback stabilizability of parabolic equations*, Appl. Math. Optim., 6 (1980), pp. 201-220.

[22] H. TRIEBEL, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, New York, Amsterdam, 1978.

[23] A. N. TYCHONOV AND A. A. SAMARSKI, *Partial Differential Equations of Mathematical Physics* (II), Holden-Day, San Francisco, 1967.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 829-847, July 1991

A PERTURBED PARALLEL DECOMPOSITION METHOD FOR A CLASS OF NONSMOOTH CONVEX MINIMIZATION PROBLEMS*

KHALIL MOUALLIF†, VAN HIEN NGUYEN‡, AND JEAN-JACQUES STRODIOT‡

Abstract. A perturbed parallel decomposition method for solving the following model problem is presented: minimize $f_0(x)+\sum_{i=1}^m f_i(x)$ over all x in $\mathbb{R}^n$, where f_0 is differentiable and strongly convex, and the f_i's are closed, proper, and convex (but not necessarily differentiable). An important feature of the proposed method is the ability to split the given problem into independent subproblems that contain only individual functions; in that way the algorithm is suitable for parallel computation. The approach also allows for inaccuracies and duality gaps during the iterative process. Furthermore, convergence is established under perturbations on the f_i's in the epiconvergence sense (and under inexact minimizations at each iteration). Relations of the method with the recent work of Han and Lou on the same model problem are also discussed. From the theoretical point of view, this paper represents a significant improvement on the parallel decomposition algorithms of Han and Lou.

Key words. decomposition technique, parallel optimization algorithm, nonsmooth convex programming, variational convergence

AMS(MOS) subject classifications. 65K05, 65K10, 90C25, 90C30

1. Introduction.

This paper presents a general perturbed parallel decomposition method for solving the following nondifferentiable convex optimization problem:

$$\text{(P)} \qquad \text{minimize}\, f_0(x)+\sum_{i=1}^{m} f_i(x) \quad \text{over all } x\in\mathbb{R}^n$$

where $f_0:\mathbb{R}^n\to\mathbb{R}$ is differentiable and strongly convex with modulus α (see, for example, Rockafellar [23]) and for all $i=1,\cdots,m$, each f_i is a closed proper convex function from $\mathbb{R}^n$ into the extended real line $\mathbb{R}\cup\{+\infty\}$ (see Rockafellar [22]). We do not assume the f_i's to be differentiable. Note that the f_i's are allowed to take the value $+\infty$; this fact is important since it permits us to model implicitly the constraints that may be imposed on the choice of x. For consistency, we assume that

$$\bigcap_{i=1}^{m} \operatorname{dom} f_i \neq \emptyset,$$

where $\operatorname{dom} f_i$ is the effective domain of f_i defined by

$$\operatorname{dom} f_i=\{x\in\mathbb{R}^n \mid f_i(x)<+\infty\}.$$

In the method we propose here, the problem (P) is first perturbed (that is, the f_i's are replaced by functions f_i^k, which approximate the f_i's in the sense of epiconvergence; see, for example, Attouch and Wets [3], [5]) and is then dualized in the sense of convex analysis (see [22]) to obtain a problem that is a nonsmooth unconstrained convex problem, which in turn is solved by applying the so-called auxiliary problem principle (see, for example, Cohen [10]). An important feature of our method is that the problem (P) (or rather, its perturbed form) is decomposed into independent subproblems that contain only individual functions f_i (or rather f_i^k); in this way the algorithm is suitable for parallel computation. In our approach we also allow the solutions of the perturbed "primal" and "dual" programs, which are alternately and

* Received by the editors January 3, 1989; accepted for publication (in revised form) September 5, 1990.

† Department of Basic Sciences, Ecole Hassania des Travaux Publics, Casablanca, Morocco.

‡ Department of Mathematics, Facultés Universitaires Notre-Dame de la Paix, B-5000 Namur, Belgium.

repeatedly solved during the parallel iterative process, to be inaccurate. Note that this property is very important, especially as far as applications and computations are concerned; exact solutions—if they do exist—may be difficult, if not impossible, to find. In the general framework that we work with in this paper (that is, under perturbations and with inexact minimizations) we are able to show that, under fairly mild conditions, the method has attractive convergence properties both in the "primal" space and in the "dual" space.

The motivation for this paper comes from the observation that the model problem (P) represents a fairly broad class of problems that may arise in various situations. For example, the present method can be applied to the following constrained optimization problem (although we are mainly concerned with the unconstrained problem (P)):

$$(\bar{\mathrm{P}}) \qquad \begin{aligned} &\text{minimize} && f_0(x) \\ &\text{subject to} && x \in C_1 \cap \cdots \cap C_m, \end{aligned}$$

where f_0 is as above, and for all $i = 1, \cdots, m$, each C_i is a nonempty closed convex set of $\mathbb{R}^n$. For consistency, we assume that the set $C_1 \cap \cdots \cap C_m$ is nonempty. Then, the problem $(\bar{\mathrm{P}})$ may be rewritten in the equivalent form

$$(1.1) \qquad \text{minimize} f_0(x) + \sum_{i=1}^{m} \delta(x|C_i) \quad \text{over all } x \in \mathbb{R}^n,$$

where $\delta(\cdot|C_i)$ is the indicator function of C_i defined by

$$\delta(x|C_i) = \begin{cases} 0 & \text{if } x \in C_i, \\ +\infty & \text{otherwise.} \end{cases}$$

Clearly, the problem (1.1) (therefore, the problem $(\bar{\mathrm{P}})$) is a problem of the form (P) since the functions $\delta(\cdot|C_i) = f_i$ are obviously closed, proper, and convex. Note that this example $(\bar{\mathrm{P}})$ includes as a special case the well-known projection problem of a given point in $\mathbb{R}^n$ onto the intersection of a finite number of convex sets of $\mathbb{R}^n$ (see [14]).

Another instance where the model problem (P) arises is when we want to

$$(1.2) \qquad \begin{aligned} &\text{minimize} && f_0(x) + \sum_{i=1}^{m} \langle c^i, z^i \rangle \\ &\text{subject to} && A_i x + B_i z^i \le b^i \quad \text{for } i = 1, \cdots, m \\ & && x \in \mathbb{R}^n, \quad z^i \in \mathbb{R}^{n_i}, \end{aligned}$$

where f_0 is as above, $c^i \in \mathbb{R}^{n_i}$, $b^i \in \mathbb{R}^{m_i}$, A_i, and B_i are $m_i \times n$ and $m_i \times n_i$ matrices, respectively, whereas $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product in $\mathbb{R}^{n_i}$ (for all $i = 1, \cdots, m$). Then, it is known (see, for example, Lasdon [16]) that, under reasonable assumptions, the functions

$$f_i(x) = \min \{\langle c^i, z^i \rangle | z^i \in \mathbb{R}^{n_i}, B_i z^i \le b^i - A_i x\}$$

are finite real-valued, convex, and piecewise linear (polyhedral). Then, the problem (1.2) can be restated as the model problem (P). Similar ideas are also encountered for other decomposable problems in stochastic programming (see, for example, the state-of-the-art tutorials of Wets [27] and Varaiya and Wets [26]) and in structural optimization (see, for example, Barthelemy [8] for a useful survey).

Our method is closely related to—but is different from—that proposed by Han and Lou [13] for the model problem (P) ("Algorithm II for (B)"). The main differences with this recent work are twofold. First of all, the derivation of the method is entirely different here from that described in [13]. Roughly speaking, in [13] Han and Lou begin by presenting a basic decomposition method for solving the following optimization problem ("Algorithm I for (A)"):

$$\text{(A)} \qquad \text{minimize}\, p(y_1, \cdots, y_m) + \sum_{i=1}^{m} g_i(y_i) \quad \text{over all } y_1 \in \mathbb{R}^{n_1}, \cdots, y_m \in \mathbb{R}^{n_m},$$

where $p:\mathbb{R}^{n_1+\cdots+n_m} \to \mathbb{R}$ is a convex, differentiable function whose gradient mapping is Lipschitz continuous, and $g_i:\mathbb{R}^{n_i} \to \mathbb{R} \cup \{+\infty\}$ is closed, proper, and convex for all $i = 1, \cdots, m$. Then, through duality theory, they transform the model problem (P) into a dual problem that has the form of (A), and apply their basic decomposition technique to the dual so that they can also decompose the model problem (P) and solve it in parallel.

A second difference between our approach and the one of Han and Lou [13] is that we allow for inaccuracies and variational perturbations during the parallel iterative process. Our development therefore encompasses that of Han and Lou for the case of the model problem (P). One of the implications of our development is that an additional regularity assumption made in [13] is proven to be superfluous. This arises from the fact that they did need a duality theorem for the dual programs they had to consider. As a consequence, our convergence results are more general. Details are discussed in § 4.

Note that our method is also related to the proximal point algorithm in convex programming (for a survey, see Lemaire [17]), but with a simultaneous utilization of a gradient method for the differentiable part of the objective function.

The remainder of the paper is organized as follows. The method is derived and stated in § 2. Its connections with Han and Lou's "Algorithm II for (B)" [13] are established in § 3. There we show that the "Algorithm I for (A)" of Han and Lou [13] can also be derived in the same way as in § 2, that is, by applying the auxiliary problem principle. Convergence results for the method are presented in § 4. Finally, we end the paper with a conclusion section.

We use the following notation and terminology. $\mathbb{R}^n$ denotes the n-dimensional Euclidean space with the ordinary inner product $\langle \cdot, \cdot \rangle$ and the associated two-norm $\|\cdot\|$. All vectors are column vectors. However, for convenience, a column vector in $\mathbb{R}^{n_1+n_2}$ is denoted by (x_1, x_2), even though x_1 and x_2 are column vectors in $\mathbb{R}^{n_1}$ and $\mathbb{R}^{n_2}$, respectively.

All the definitions and most of the notation concerning convex analysis are those of Rockafellar [22]. For instance, the conjugate of a convex function f, denoted by f^*, is given by

$$f^*(y) = \sup_{x \in \mathbb{R}^n} \{\langle x, y\rangle - f(x)\}$$

and the subdifferential of f at $\bar{x}$ is given by

$$\partial f(\bar{x}) = \{y \in \mathbb{R}^n \mid f(z) \geq f(\bar{x}) + \langle z - \bar{x}, y\rangle \text{ for all } z \in \mathbb{R}^n\}.$$

2. Derivation of the method. In this section, the general perturbed parallel decomposition method will be proposed for solving the model problem (P). Before doing this it is necessary to recall (see Wijsman [29]) the following definition of epiconvergence of a sequence of closed proper convex functions. Epigraphical analysis provides a very rich and unified tool to study a broad class of problems that includes variational problems, generalized equations, and mathematical programs, and has received much attention during the last decade (see, for example, Attouch [2], Attouch and Wets [3], [5], and Rockafellar and Wets [24]).

Let $\{\phi^k\}$ be a sequence of closed proper convex functions from $\mathbb{R}^n$ into $\mathbb{R} \cup \{+\infty\}$ and suppose ϕ is a function satisfying the same hypotheses. Epiconvergence of $\{\phi^k\}$ to ϕ may be characterized by the conjunction of these very tangible local conditions: At each $x \in \mathbb{R}^n$,

(i) Whenever $\{x^k\} \to x$, then $\liminf \phi^k(x^k) \geq \phi(x)$;

(ii) There exists a sequence $\{x^k\}$ convergent to x for which $\lim \phi^k(x^k) = \phi(x)$. We will write

$$\phi = \text{epi-lim } \phi^k$$

for "$\{\phi^k\}$ is epi-convergent to ϕ."

As in Lemaire [18], the model problem (P) is embedded in a family of parametrized problems by perturbing the functions f_i (for $i = 1, \cdots, m$) in replacing them by the functions f_i^k (for $i = 1, \cdots, m$, and $k \in \mathbb{N}$) satisfying the following condition:

$$f_i = \underset{k \to +\infty}{\text{epi-lim}} f_i^k \qquad (i = 1, \cdots, m), \tag{2.1}$$

where, of course, the functions f_i^k are also closed, proper, and convex. Thus, we are led to consider the approximating problems

$$\text{minimize } f_0(x) + \sum_{i=1}^{m} f_i^k(x) \quad \text{over all } x \in \mathbb{R}^n. \tag{P$_k$}$$

For illustration purposes, we consider the problem ($\bar{\text{P}}$) here again. It was shown in § 1 that the problem ($\bar{\text{P}}$) is equivalent to the problem (1.1), which is of the form (P) with

$$f_i = \delta(\cdot | C_i) \quad \text{for } i = 1, \cdots, m.$$

For each $i = 1, \cdots, m$, we consider penalty functions f_i^k (where $k \in \mathbb{N}$) satisfying the following conditions (compare with Auslender, Crouzeix, and Fédit [7]):

$$f_i^k : \mathbb{R}^n \to \mathbb{R} \quad \text{is convex for all } k \in \mathbb{N}; \tag{H1}$$

$$f_i^k \leq f_i^{k+1} \quad \text{for all } k \in \mathbb{N}; \tag{H2}$$

$$\begin{aligned} &\lim_{k \to +\infty} f_i^k(x) = 0 && \text{if } x \in C_i, \\ &\lim_{k \to +\infty} f_i^k(x) = +\infty && \text{otherwise.} \end{aligned} \tag{H3}$$

Then, it follows immediately from Attouch [2, Thm. 3.20] that

$$\underset{k \to +\infty}{\text{epi-lim}} f_i^k = \delta(\cdot | C_i) = f_i \quad \text{for all } i = 1, \cdots, m.$$

Note that if the constraint set C_i, for $i = 1, \cdots, m$ is defined by

$$C_i = \{x \in \mathbb{R}^n | g_i(x) \leq 0\},$$

where $g_i : \mathbb{R}^n \to \mathbb{R}$ is convex, then we can take f_i^k as the classical exterior penalty function

$$f_i^k = r_i^k(\max\{0, g_i\})^2$$

or as the exact penalty function

$$f_i^k = r_i^k \max\{0, g_i\},$$

where

$$0 < r_i^k \leq r_i^{k+1} \quad \text{and} \quad \lim_{k \to +\infty} r_i^k = +\infty.$$

In these conditions, it is easy to see that the functions f_i^k satisfy the above assumptions (H1)-(H3); that is, they are perturbations of f_i in our sense.

For discussion of other examples of f_i^k verifying (2.1), we refer to Attouch and Wets [3], Mine and Fukushima [21], Attouch [2], and Alart and Lemaire [1].

We now turn to the perturbed problem (P_k). We take the Fenchel dual (see [22, Thm. 31.1]) of (P_k) to obtain a nonsmooth dual convex problem, namely,

$$(\mathrm{P}_k^*) \qquad \text{minimize} f_0^*(y)+\left(\sum_{i=1}^m f_i^k\right)^*(-y) \quad \text{over all } y\in\mathbb{R}^n.$$

By Theorem 16.4 of Rockafellar [22], if

$$(2.2) \qquad \bigcap_{i=1}^m \operatorname{ri}(\operatorname{dom} f_i^k)\neq\varnothing$$

(ri stands for the relative interior) then the second term of the objective function of (P_k^*) is such that

$$\left(\sum_{i=1}^m f_i^k\right)^*(y)=\inf\left\{\sum_{i=1}^m (f_i^k)^*(y_i)|y_1,\cdots,y_m\in\mathbb{R}^n,\ \sum_{i=1}^m y_i=y\right\}.$$

Because of this equality, we consider from now on (even in the cases where the hypothesis (2.2) does not hold) as the Fenchel dual program of (P_k) the following problem:

$$(\mathrm{D}_k) \qquad \text{minimize} f_0^*\left(\sum_{i=1}^m y_i\right)+\sum_{i=1}^m (f_i^k)^*(-y_i) \quad \text{over all } y_1\in\mathbb{R}^n,\cdots,y_m\in\mathbb{R}^n.$$

It should be pointed out that, in this paper, we *do not* assume that the regularity condition (2.2) (or a similar condition on ri (dom f_i)) holds. As a consequence of this, we do not necessarily have a duality theorem for the dual pair (P_k) and (D_k). Another way of expressing this is to say that our approach allows for duality gaps between these primal and dual programs during the parallel iterative process. This property is important because it may not always be possible or desirable to generate dual solutions that close the duality gap.

The following result is the key to the derivation of the method. To state it some new terminology must first be introduced.

Let $F_0:\mathbb{R}^N\to\mathbb{R}$ be a differentiable convex function and let $G^k:\mathbb{R}^N\to\mathbb{R}\cup\{+\infty\}$ be a closed proper convex function depending on a natural integer k. We are concerned with the following (master) problem for each fixed k in $\mathbb{N}$:

$$(\mathrm{MP}_k) \qquad \text{minimize } F_0(y)+G^k(y) \quad \text{over all } y\in\mathbb{R}^N.$$

Now let $K^k:\mathbb{R}^n\to\mathbb{R}$ be a differentiable and strongly convex function depending on k and consider the following function $J_{\bar{y}}^k:\mathbb{R}^N\to\mathbb{R}\cup\{+\infty\}$ depending on some $\bar{y}\in\mathbb{R}^N$ and $\varepsilon_k>0$ defined by

$$(2.3) \qquad J_{\bar{y}}^k(y)=K^k(y)+\langle\varepsilon_k\nabla F_0(\bar{y})-\nabla K^k(\bar{y}),y\rangle+\varepsilon_k G^k(y) \quad \text{for all } y\in\mathbb{R}^N.$$

Then we have the following fundamental lemma that Cohen [10, Lemma 2.2] called the auxiliary problem principle.

LEMMA 1. *For each fixed $k\in\mathbb{N}$, we have that $\bar{y}\in\mathbb{R}^N$ solves the (auxiliary) problem*

$$(\mathrm{AP}_k) \qquad \textit{minimize } J_{\bar{y}}^k(y) \quad \textit{over all } y\in\mathbb{R}^N$$

if and only if $\bar{y}$ solves the master problem (MP_k).

Proof. The following statements are equivalent:

$\bar{y}$ is an optimal solution of $(\mathrm{AP}_k) \Leftrightarrow 0 \in \partial J_{\bar{y}}^k(\bar{y})$

$\Leftrightarrow 0 \in \nabla F_0(\bar{y}) + \partial G^k(\bar{y})$ by [22, Thm. 23.8])

$\Leftrightarrow \bar{y}$ is an optimal solution of (MP_k). □

To find such a $\bar{y} \in \mathbb{R}^N$ (whose existence is ensured under reasonable conditions) Cohen [10] proposed to use, for each fixed $k \in \mathbb{N}$, a kind of fixed-point algorithm.

Let a sequence of differentiable and strongly convex functions $\{K^{kj}\}_{j \in \mathbb{N}}$ and a sequence of positive numbers $\{\varepsilon_{kj}\}_{j \in \mathbb{N}}$ be chosen. The index j denotes the iteration index. Cohen's algorithm [10] can be described as follows:

Initialization Step
Choose a point $y^0 \in \mathbb{R}^N$, let $j = 0$, and go to the main step.
Main Step
Let y^{j+1} be an optimal solution to the following auxiliary problem:

(AP_{kj}) minimize $K^{kj}(y) + \langle \varepsilon_{kj} \nabla F_0(y^j) - \nabla K^{kj}(y^j), y \rangle + \varepsilon_{kj} G^k(y)$ for all $y \in \mathbb{R}^N$.

If $\|y^{j+1} - y^j\|$ or $|F_0(y^{j+1}) + G^k(y^{j+1}) - F_0(y^j) - G^k(y^j)|$ is below some desired tolerance, then stop; otherwise, replace j by $j+1$, and repeat.

Set for each fixed $k \in \mathbb{N}$

$$F_0(y) = f_0^*\left(\sum_{i=1}^m y_i\right) \tag{2.4}$$

and

$$G^k(y) = \sum_{i=1}^m (f_i^k)^*(-y_i) \tag{2.5}$$

for all $y = (y_1, \ldots, y_m) \in \mathbb{R}^{m \cdot n}$. Also we set $m \cdot n = N$. Then, clearly, the Fenchel dual problem (D_k) is of the form of the master problem (MP_k), to which Lemma 1 applies. Indeed, because $f_0 : \mathbb{R}^n \to \mathbb{R}$ was supposed to be differentiable and strongly convex with modulus α, it is known (see, for example, Han and Lou [12, § 2] or [13, § 3]) that the conjugate function $f_0^* : \mathbb{R}^n \to \mathbb{R}$ of f_0 is also differentiable everywhere, strictly convex, co-finite, and

$$\nabla f_0^* = (\nabla f_0)^{-1}. \tag{2.6}$$

Thus, the function $F_0 : \mathbb{R}^N \to \mathbb{R}$, as defined by (2.4), is differentiable and convex. On the other hand, because the functions $f_i^k : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ were assumed to be closed, proper, and convex, it is easy to see (by Rockafellar [22, Thm. 12.2]) that the conjugate functions $(f_i^k)^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ of f_i^k also have the same properties. For each $k \in \mathbb{N}$, the function $G^k : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$, as defined by (2.5), is thus obviously a closed proper convex function.

However, for obvious reasons, we do not want to solve each dual problem (D_k). So it is natural to consider a diagonal process to generate the sequence $\{y^k\}_{k \in \mathbb{N}}$: At the kth iteration, in order to compute y^{k+1} from y^k, we only solve the auxiliary problem (AP_{kj}) with $j = k$. Moreover, for each $k \in \mathbb{N}$, we suppose that $\varepsilon_{kk} = 1$ and $K^{kk}(y) = (1/2\lambda_k)\|y\|^2$ for all $y \in \mathbb{R}^N$, where $\lambda_k > 0$. This kind of quadratic core K^{kk} has often been used to ensure convergence; in particular, it is akin to the proximal point algorithm; for example, see Rockafellar [23] and Lemaire [17].

In other words, from a starting point $y^0 \in \mathbb{R}^N$, a sequence $\{y^k\}_{k\in\mathbb{N}}$ will be generated step by step as follows: for each $k \in \mathbb{N}$, y^{k+1} is the optimal solution to the problem

$$(\mathrm{AD}_k) \quad \text{minimize } (1/2\lambda_k)\|y\|^2 + \langle \nabla F_0(y^k) - (1/\lambda_k)y^k, y\rangle + G^k(y) \quad \text{for all } y \in \mathbb{R}^N.$$

Then we can write

$$y^{k+1} = \arg\min_{y\in\mathbb{R}^N} \{(1/2\lambda_k)\|y - y^k + \lambda_k \nabla F_0(y^k)\|^2 + G^k(y)\} \tag{2.7}$$

because

$$(1/2\lambda_k)\|y - y^k + \lambda_k \nabla F_0(y^k)\|^2 + G^k(y) = (1/2\lambda_k)\|y\|^2 + \langle \nabla F_0(y^k) - (1/\lambda_k)y^k, y\rangle + G^k(y) + \text{term independent of } y.$$

Now remember from (2.4) and (2.5) that

$$\nabla F_0(y) = \nabla F_0(y_1, \cdots, y_m) = \left(\nabla f_0^*\left(\sum_{i=1}^m y_i\right), \cdots, \nabla f_0^*\left(\sum_{i=1}^m y_i\right)\right)$$

and

$$G^k(y) = \sum_{i=1}^m (f_i^k)^*(-y_i),$$

and that $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^{m\cdot n}$, in such a way that we can get the following splitting of y^{k+1}, as defined by (2.7), into individual components ($i = 1, \cdots, m$):

$$y_i^{k+1} = \arg\min_{y_i\in\mathbb{R}^n} \left\{(1/2\lambda_k)\left\| y_i - y_i^k + \lambda_k \nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)\right\|^2 + (f_i^k)^*(-y_i)\right\}.$$

Thus, we easily obtain that

$$y_i^{k+1} = \arg\min_{y_i\in\mathbb{R}^n} \left\{(1/2\lambda_k)\|y_i - y_i^k\|^2 + \left\langle y_i - y_i^k, \nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)\right\rangle + (f_i^k)^*(-y_i)\right\}. \tag{2.8}$$

Consequently, the characterization of optimality says that

$$0 \in \partial\left((1/2\lambda_k)\|\cdot - y_i^k\|^2 + \left\langle \cdot - y_i^k, \nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)\right\rangle + (f_i^k)^*(-\cdot)\right)(y_i^{k+1}).$$

Convex analysis calculus then gives

$$0 \in (1/\lambda_k)(y_i^{k+1} - y_i^k) + \nabla f_0^*\left(\sum_{i=1}^m y_i^k\right) - \partial(f_i^k)^*(-y_i^{k+1}).$$

Equivalently, we have

$$(1/\lambda_k)(y_i^{k+1} - y_i^k) + \nabla f_0^*\left(\sum_{i=1}^m y_i^k\right) \in \partial(f_i^k)^*(-y_i^{k+1}).$$

By Rockafellar [22, Thm. 23.5], this is in turn equivalent to

$$-y_i^{k+1} \in \partial f_i^k\left((1/\lambda_k)(y_i^{k+1} - y_i^k) + \nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)\right);$$

that is,

$$0 \in y_i^{k+1} + \partial f_i^k\left((1/\lambda_k)(y_i^{k+1} - y_i^k) + \nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)\right).$$

We can then express this differential inclusion as

$$y_i^{k+1} = \arg\min_{y_i \in \mathbb{R}^n} \left\{(1/2\lambda_k)\|y_i\|^2 + f_i^k\left((1/\lambda_k)(y_i - y_i^k) + \nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)\right)\right\}. \tag{2.9}$$

To evaluate $\nabla f_0^*(\sum_{i=1}^m y_i^k)$ in (2.9), we want to avoid the conjugate function f_0^* in computation. For this purpose, we may calculate $\nabla f_0^*(\sum_{i=1}^m y_i^k)$ by solving the following problem:

$$\text{minimize } f_0(x) - \left\langle x, \sum_{i=1}^m y_i^k \right\rangle \quad \text{over all } x \in \mathbb{R}^n. \tag{F}$$

To see this, let x^k be a solution of (F). Then we have the following equivalence:

$$\nabla f_0(x^k) - \sum_{i=1}^m y_i^k = 0 \Leftrightarrow x^k = (\nabla f_0)^{-1}\left(\sum_{i=1}^m y_i^k\right).$$

From (2.6), we obtain

$$x^k = \nabla f_0^*\left(\sum_{i=1}^m y_i^k\right), \tag{2.10}$$

which was what we wanted.

Until now, we have been dealing with the method in its perturbed form involving exact minimizations. Now we will take inexact minimizations into consideration. To this end, let us introduce $m+1$ sequences of nonnegative numbers: $\{\varepsilon_i^k\}_{k\in\mathbb{N}}$ for $i = 0, 1, \cdots, m$. Also let $\{\lambda_k\}_{k\in\mathbb{N}}$ be a sequence of positive numbers.

To make our notation simpler, we set

$$(1/\lambda_k)(y_i - y_i^k) + \nabla f_0^*\left(\sum_{i=1}^m y_i^k\right) = z_i.$$

This and (2.10) give that (2.9) is equivalent to

$$\begin{aligned} z_i^{k+1} &= \arg\min_{z_i \in \mathbb{R}^n} \{(1/2\lambda_k)\|y_i^k + \lambda_k(z_i - x^k)\|^2 + f_i^k(z_i)\}, \\ y_i^{k+1} &= y_i^k + \lambda_k(z_i^{k+1} - x^k). \end{aligned} \tag{2.11}$$

Taking all these into account, we are now in a position to state our general perturbed parallel decomposition method as follows.

THE METHOD (in its general form). Let $\lambda_k > 0$ and $\varepsilon_i^k \geq 0$ for $i = 0, 1, \cdots, m$ and $k = 0, 1, 2, \cdots$. We start with any point $y^0 = (y_1^0, \cdots, y_m^0) \in \mathbb{R}^{m\cdot n}$. We then calculate $x^0 \in \mathbb{R}^n$, then $y^1 = (y_1^1, \cdots, y_m^1) \in \mathbb{R}^{m\cdot n}$, $x^1 \in \mathbb{R}^n, \cdots$ by doing the following computation:

(a) Compute x^k by

$$x^k \in \varepsilon_0^k - \arg\min_{x \in \mathbb{R}^n} \left\{f_0(x) - \left\langle x, \sum_{i=1}^m y_i^k \right\rangle\right\}. \tag{2.12}$$

(b) For $i = 1, \cdots, m$, we compute y_i^{k+1} by

$$\begin{aligned} z_i^{k+1} &\in \varepsilon_i^k - \arg\min_{z_i \in \mathbb{R}^n} \{(1/2\lambda_k)\|y_i^k + \lambda_k(z_i - x^k)\|^2 + f_i^k(z_i)\}, \\ y_i^{k+1} &= y_i^k + \lambda_k(z_i^{k+1} - x^k). \end{aligned} \tag{2.13}$$

As usual we let

$$\varepsilon - \arg\min_{x\in\mathbb{R}^n} \phi(x) = \{\bar{x}\in\mathbb{R}^n | \phi(\bar{x}) \le \inf \phi + \varepsilon\}$$

for any function $\phi:\mathbb{R}^n \to \mathbb{R}\cup\{+\infty\}$.

At this point, a few remarks are of interest. Steps (a) *and* (b) of the method are realized with only approximate minimization in the subproblems. The subproblem in Step (a) contains only the (differentiable strongly convex) function f_0. Each of the m subproblems in step (b) contains only one (nondifferentiable convex) function f_i^k. These m subproblems in step (b) are independent of each other and can be solved in parallel by using any nonsmooth convex minimization algorithm (see especially Auslender [6]; see also Fukushima [11], Kiwiel [15], Strodiot and Nguyen [25] and the references cited therein). Therefore, in the proposed method, the parallel decomposition takes place with respect to the set of "constraints" f_i or rather f_i^k (as opposed to with respect to the set of varaibles x). Note that the objective function of each subproblem in step (b) (which is defined by an approximation f_i^k of f_i) is not only approximately solved but also depends on an approximate solution to the subproblem in step (a).

3. Connections with the work of Han and Lou. In [13] Han and Lou give a parallel decomposition algorithm to solve the model problem (P). Because in the present paper we are concerned with the same problem (P) (with the same assumptions on the problem functions f_0 and f_i ($i=1,\cdots,m$)), it is natural to establish here the relationship between our method and the one proposed by Han and Lou. In order to facilitate the understanding of the analysis of this section we first briefly recall Han and Lou's algorithm ("Algorithm II for (B)") in [13].

HAN AND LOU'S ALGORITHM. Start from a point $y^0=(y_1^0,\cdots,y_m^0)$ in $\mathbb{R}^{m\cdot n}$ and a sufficiently large number a (>0). Having $y^k=(y_1^k,\cdots,y_m^k)$ at the kth iteration, we do the following:

(a) Compute x^k by solving the problem

$$\text{(3.1)}\qquad \text{minimize}\, f_0(x) - \left\langle x, \sum_{i=1}^{m} y_i^k \right\rangle \quad \text{over all } x\in\mathbb{R}^n.$$

(b) For $i=1,\cdots,m$, we solve the problem

$$\text{(3.2)}\qquad \text{minimize}\, a\|y_i\|^2 + f_i(2ay_i - 2ay_i^k + x^k) \quad \text{over all } y_i\in\mathbb{R}^n$$

to obtain the optimal solution y_i^{k+1} (ith component of $y^{k+1}=(y_1^{k+1},\cdots,y_m^{k+1})$).

Now, the relationship we just mentioned above can be laid out as in the next result.

PROPOSITION 1. *For the model problem* (P), *Han and Lou's algorithm as defined by* (3.1)-(3.2) *is a particular case of our method as defined by* (2.12)-(2.13) *with the following special choices*:

$$f_i^k = f_i \quad \textit{for all } i=1,\cdots,m \textit{ and all } k\in\mathbb{N},$$

$$\varepsilon_i^k = 0 \textit{ for all } i=0,1,\cdots,m \textit{ and all } k\in\mathbb{N}.$$

Proof. The proof is immediate from the derivation of our method in the previous section (see (2.9), (2.10), and (F)) in taking $1/2\lambda_k = a$ for all $k\in\mathbb{N}$. □

On the other hand, recall that in [13], the authors started with a basic decomposition method optimization problem they called problem (A) ("Algorithm I for (A)"); see § 1 for the definition of (A). The algorithm that they proposed for solving this problem (A) (and from which they derived, through duality, their "Algorithm II for (B)") can be stated as follows.

Start from a point $y^0=(y_1^0,\cdots,y_m^0)$ in $\mathbb{R}^{n_1+\cdots+n_m}$ and a real number $a>0$. At the kth iteration, having $y^k=(y_1^k,\cdots,y_m^k)\in\mathbb{R}^{n_1+\cdots+n_m}$, we compute $y^{k+1}=(y_1^{k+1},\ldots,y_m^{k+1})\in\mathbb{R}^{n_1+\cdots+n_m}$ in doing the following:

(a) Calculate

$$z^k=(z_1^k,\cdots,z_m^k)=\nabla p(y^k)\in\mathbb{R}^{n_1+\cdots+n_m}. \tag{3.3}$$

(b) For $i=1,\cdots,m$, find $y_i^{k+1}\in\mathbb{R}^{n_i}$ by solving the problem

$$\text{minimize } a\|y_i-y_i^k\|^2+\langle y_i-y_i^k, z_i^k\rangle+g_i(y_i) \quad \text{over all } y_i\in\mathbb{R}^{n_i}. \tag{3.4}$$

Then, it is not difficult to see that Han and Lou's "Algorithm I for (A)" as defined by (3.3)-(3.4) can also be viewed as a particular case of our method as defined by (2.7), or equivalently (2.8). Indeed, it suffices to take

$$F_0(y)=p(y),\quad G^k(y)=\sum_{i=1}^m g_i(y_i),\quad y=(y_1,\cdots,y_m)\in\mathbb{R}^{n_1+\cdots+n_m}=\mathbb{R}^N$$

and

$$\frac{1}{2\lambda_k}=a \quad \text{for all } k\in\mathbb{N}.$$

We conclude this section with the following remark. As Han and Lou [13, § 4] point out, their "Algorithm II for (B)" is the one considered in [12] for the case of minimizing the constrained problem ($\bar{\text{P}}$). Consequently, our method may also be viewed as generalizing the first parallel algorithm of Han and Lou [12] to solve ($\bar{\text{P}}$).

4. Convergence of the general method. This section presents our analysis of convergence for the general method as defined by (2.12)-(2.13) to solve the model problem (P). We begin by studying an estimate on the distance between the solution and an epsilon-solution for a problem of minimizing a strongly convex function (see also Auslender [6] and Auslender, Crouzeix, and Fédit [7, Lemma 2.2]).

LEMMA 2. *Let $\phi:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ be a closed proper strongly convex function (with modulus $\beta>0$) and let $\varepsilon>0$. Let $\bar{x}$ be the exact solution to the problem of minimizing ϕ over $\mathbb{R}^n$ and suppose that $\bar{x}_\varepsilon$ is an ε-minimizer of the same optimization problem. Then*

$$\|\bar{x}-\bar{x}_\varepsilon\|\le\sqrt{2\varepsilon/\beta}. \tag{4.1}$$

Proof. By Proposition 6 of [23] we have

$$\phi(\bar{x}_\varepsilon)\ge\phi(\bar{x})+(\beta/2)\|\bar{x}-\bar{x}_\varepsilon\|^2.$$

On the other hand, from the definition of $\bar{x}_\varepsilon$,

$$\phi(\bar{x}_\varepsilon)\le\phi(\bar{x})+\varepsilon.$$

Combining this with the preceding inequality yields (4.1). □

For any closed proper convex function $\phi:\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ and any real number $\lambda>0$, we set

$$\phi_\lambda(x)=\inf_{z\in\mathbb{R}^n}\{\phi(z)+(1/2\lambda)\|x-z\|^2\}. \tag{4.2}$$

These functions ϕ_λ, for $\lambda>0$, play a fundamental role in optimization theory and are called Moreau-Yosida approximates (of index λ of ϕ). It is well known that $\phi_\lambda:\mathbb{R}^n\to\mathbb{R}$ is a closed convex function. The unique $\bar{z}\in\mathbb{R}^n$ for which the infimum in (4.2) is attained, is denoted by prox $(x|\lambda\phi)$ and is characterized by

$$(1/\lambda)(x-\text{prox}\,(x|\lambda\phi))\in\partial\phi(\text{prox}\,(x|\lambda\phi)). \tag{4.3}$$

This point $\operatorname{prox}(x|\lambda\phi)$ is called the proximal point, and the mapping $\operatorname{prox}(\cdot|\lambda\phi)$ the proximal mapping associated with $\lambda\phi$. It is nonexpansive, as shown below:

$$\|\operatorname{prox}(x|\lambda\phi)-\operatorname{prox}(x'|\lambda\phi)\| \le \|x-x'\| \quad \text{for all } x, x' \in \mathbb{R}^n. \tag{4.4}$$

More detailed information on this can be found in Brézis [9], Attouch [2], and Rockafellar [22], [23].

For subsequent use, define the functions g_i and g_i^k (for all $i=1,\cdots,m$ and all $k\in\mathbb{N}$) by letting

$$g_i(y_i)=f_i^*(-y_i) \quad \text{and} \quad g_i^k(y_i)=(f_i^k)^*(-y_i) \quad \text{for all } y_i\in\mathbb{R}^n. \tag{4.5}$$

The next theorem provides the first important step in obtaining results of convergence for our general method for the model problem (P).

THEOREM 1. *Under the same assumptions and with the same notation as above, there exists a sequence* $\{e_0^k\}_{k\in\mathbb{N}}$ *of points in* $\mathbb{R}^n$ *and a sequence* $\{e^k\}_{k\in\mathbb{N}}$ *of points in* $\mathbb{R}^{m\cdot n}$, *where* $e^k=(e_1^k,\cdots,e_m^k)$ *for all* $k\in\mathbb{N}$, *such that*

$$x^k=\nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)+e_0^k \quad \textit{for all } k\in\mathbb{N}, \tag{4.6}$$

$$y_i^{k+1}=\operatorname{prox}\left(y_i^k-\lambda_k\nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)\middle|\lambda_k g_i^k\right)+e_i^k \quad \textit{for all } i=1,\cdots,m \textit{ and all } k\in\mathbb{N}, \tag{4.7}$$

where the errors e_0^k *and* e_i^k *satisfy the following estimates*:

$$\|e_0^k\|\le\sqrt{2\varepsilon_0^k/\alpha} \quad \textit{for all } k\in\mathbb{N}, \tag{4.8}$$

$$\|e_i^k\|\le\sqrt{2\lambda_k\varepsilon_i^k}+\lambda_k\sqrt{2\varepsilon_0^k/\alpha} \quad \textit{for all } i=1,\cdots,m \textit{ and all } k\in\mathbb{N}. \tag{4.9}$$

Proof. It follows from (2.10) that

$$\nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)=\arg\min_{x\in\mathbb{R}^n}\left\{f_0(x)-\left\langle x, \sum_{i=1}^m y_i^k\right\rangle\right\}. \tag{4.10}$$

So, using (4.10), (2.12), and Lemma 2, we obtain

$$\left\|x^k-\nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)\right\|\le\sqrt{2\varepsilon_0^k/\alpha} \tag{4.11}$$

because the function $f_0(\cdot)-\langle\cdot,\sum_{i=1}^m y_i^k\rangle:\mathbb{R}^n\to\mathbb{R}$ is strongly convex with modulus α. Then, it suffices to set

$$e_0^k=x^k-\nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)$$

in order to get the desired results (4.6) and (4.8).

To show that we have (4.7) and (4.9) let us define, for $i=1,\cdots,m$ and $k\in\mathbb{N}$, the points $\bar{y}_i^{k+1}$ and $\tilde{y}_i^{k+1}$ as follows:

$$\bar{y}_i^{k+1}=\arg\min_{y_i\in\mathbb{R}^n}\{(1/2\lambda_k)\|y_i\|^2+f_i^k((1/\lambda_k)(y_i-y_i^k)+x^k)\} \tag{4.12}$$

and

$$\tilde{y}_i^{k+1}=\operatorname{prox}\left(y_i^k-\lambda_k\nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)\middle|\lambda_k g_i^k\right). \tag{4.13}$$

Now, recall that y_i^{k+1}, which was defined by (2.13), can be written as

$$y_i^{k+1}\in\varepsilon_i^k-\arg\min_{y_i\in\mathbb{R}^n}\{(1/2\lambda_k)\|y_i\|^2+f_i^k((1/\lambda_k)(y_i-y_i^k)+x^k)\}\tag{4.14}$$

(to see this, it suffices to set $y_i=y_i^k+\lambda_k(z_i-x^k)$).

Let

$$e_i^k=y_i^{k+1}-\tilde{y}_i^{k+1}.$$

Then we have immediately (4.7). Let us prove that we have the estimate (4.9) for e_i^k. We have

$$\|e_i^k\|\le\|y_i^{k+1}-\bar{y}_i^{k+1}\|+\|\bar{y}_i^{k+1}-\tilde{y}_i^{k+1}\|.\tag{4.15}$$

Because the function $(1/2\lambda_k)\|\cdot\|^2+f_i^k((1/\lambda_k)(\cdot-y_i^k)+x^k):\mathbb{R}^n\to\mathbb{R}\cup\{+\infty\}$ is strongly convex with modulus $1/\lambda_k$, we deduce from Lemma 2 and the definition of $\bar{y}_i^{k+1}$ (see (4.12)) and of y_i^{k+1} (see (4.14)) that

$$\|y_i^{k+1}-\bar{y}_i^{k+1}\|\le\sqrt{2\lambda_k\varepsilon_i^k}.\tag{4.16}$$

It remains to estimate the second term of the right-hand side of (4.15) in order to finish the proof of the theorem. For this purpose, let us write the following equivalent statements:

$$\begin{aligned}(4.12)&\Leftrightarrow 0\in\bar{y}_i^{k+1}+\partial f_i^k((1/\lambda_k)(\bar{y}_i^{k+1}-y_i^k)+x^k)\\&\Leftrightarrow(1/\lambda_k)(\bar{y}_i^{k+1}-y_i^k)+x^k\in\partial(f_i^k)^*(-\bar{y}_i^{k+1})\\&\Leftrightarrow 0\in(1/\lambda_k)(\bar{y}_i^{k+1}-y_i^k)+x^k-\partial(f_i^k)^*(-\bar{y}_i^{k+1})\\&\Leftrightarrow 0\in(1/\lambda_k)(\bar{y}_i^{k+1}-y_i^k)+x^k+\partial g_i^k(\bar{y}_i^{k+1})\quad\text{(see (4.5))}\\&\Leftrightarrow\bar{y}_i^{k+1}=\arg\min_{y_i\in\mathbb{R}^n}\{(1/2\lambda_k)\|y_i-y_i^k+\lambda_kx^k\|^2+g_i^k(y_i)\}.\end{aligned}$$

That is, we have

$$\bar{y}_i^{k+1}=\operatorname{prox}(y_i^k-\lambda_kx^k|\lambda_kg_i^k).\tag{4.17}$$

Because the proximal mapping is nonexpansive (see (4.4)), it follows from (4.13), (4.17), and (4.11) that

$$\begin{aligned}\|\bar{y}_i^{k+1}-\tilde{y}_i^{k+1}\|&\le\lambda_k\left\|x^k-\nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)\right\|\\&\le\lambda_k\sqrt{2\varepsilon_0^k/\alpha}.\end{aligned}$$

This inequality together with (4.16) and (4.15) gives the desired result (4.9), and the proof is finished. □

Before stating the next corollary, we recall that the functions F_0 and G^k, defined for all $y=(y_1,\cdots,y_m)\in\mathbb{R}^{m\cdot n}$, were introduced in § 2 by (2.4) and (2.5), respectively (see also (4.5)).

COROLLARY 1. *For all* $k\in\mathbb{N}$, *we have*

$$y^{k+1}=\operatorname{prox}(y^k-\lambda_k\nabla F_0(y^k)|\lambda_kG^k)+e^k,\tag{4.18}$$

where the error e^k *satisfies the following estimate*:

$$\|e^k\|\le\sqrt{m}\,(\sqrt{2\lambda_k\varepsilon_k}+\lambda_k\sqrt{2\varepsilon_0^k/\alpha})\tag{4.19}$$

where $\varepsilon_k=\max\{\varepsilon_1^k,\cdots,\varepsilon_m^k\}$.

To prove Corollary 1 the following lemma is used.

LEMMA 3. *For each $i=1,\cdots,m$, let $\phi_i:\mathbb{R}^{n_i}\to\mathbb{R}\cup\{+\infty\}$ be a closed proper convex function. For $N=n_1+\cdots+n_m$ define the function $\Phi:\mathbb{R}^N\to\mathbb{R}\cup\{+\infty\}$ by letting*

$$\Phi(y)=\sum_{i=1}^m \phi_i(y_i) \quad \textit{for all } y=(y_1,\cdots,y_m)\in\mathbb{R}^N. \tag{4.20}$$

Then Φ is a closed proper convex function, and we have, for all $\lambda>0$,

$$\Phi_\lambda(y)=\sum_{i=1}^m (\phi_i)_\lambda(y_i) \tag{4.21}$$

and

$$\operatorname{prox}(y|\lambda\Phi)=(\operatorname{prox}(y_1|\lambda\phi_1),\cdots,\operatorname{prox}(y_m|\lambda\phi_m)). \tag{4.22}$$

Proof. The proof of the fact that Φ is closed, proper, and convex is omitted here because it is easy. To prove (4.21) note that, for $\lambda>0$, we successively have

$$\begin{aligned}\Phi_\lambda(y)&=\inf\{(1/2\lambda)\|z-y\|^2+\Phi(z)|z=(z_1,\cdots,z_m)\in\mathbb{R}^N\}\\&=\inf\left\{\sum_{i=1}^m[(1/2\lambda)\|z_i-y_i\|^2+\phi_i(z_i)]|z\in\mathbb{R}^N\right\}\\&=\sum_{i=1}^m\inf\{(1/2\lambda)\|z_i-y_i\|^2+\phi_i(z_i)|z_i\in\mathbb{R}^{n_i}\}\\&=\sum_{i=1}^m(\phi_i)_\lambda(y_i),\end{aligned}$$

where $(\phi_i)_\lambda$ denotes the Moreau-Yosida approximate of index λ of ϕ_i.

On the other hand, to prove (4.22), let us set

$$z=\operatorname{prox}(y|\lambda\Phi)=(z_1,\cdots,z_m)\in\mathbb{R}^N.$$

Because of the characterization (4.3) of the proximal point, we have

$$(1/\lambda)(y-z)\in\partial\Phi(z).$$

By virtue of Lemma 4.3 of McLinden [19], this inclusion is equivalent to

$$(1/\lambda)(y_i-z_i)\in\partial\phi_i(z_i)\quad\text{for all } i=1,\cdots,m;$$

this means that

$$z_i=\operatorname{prox}(y_i|\lambda\phi_i)\quad\text{for all } i=1,\cdots,m,$$

and the lemma is proved. □

Proof of Corollary 1. The formulation (4.18) is a direct consequence of (4.22), (4.7), and the fact that

$$\nabla F_0(y^k)=\nabla F_0(y_1^k,\cdots,y_m^k)=\left(\nabla f_0^*\left(\sum_{i=1}^m y_i^k\right),\cdots,\nabla f_0^*\left(\sum_{i=1}^m y_i^k\right)\right).$$

On the other hand, the estimate (4.19) is immediate from (4.9). □

We now give a first convergence theorem of the general method. To state this we introduce the following Fenchel dual program of (P):

$$\text{(D)}\qquad \text{minimize} f_0^*\left(\sum_{i=1}^m y_i\right)+\sum_{i=1}^m f_i^*(-y_i)\quad\text{over all } y_1\in\mathbb{R}^n,\cdots,y_m\in\mathbb{R}^n.$$

THEOREM 2. *Consider the general method as defined by* (2.12)–(2.13). *Let* $\{x^k\}_{k\geq 0}$ *and* $\{y^k\}_{k\geq 1}$ *be the sequences it generates from an arbitrary initial point* $y^0 \in \mathbb{R}^{m\cdot n}$. *Suppose the following*:

(a) $0<\underline{\lambda}\leq \lambda_k \leq \bar{\lambda} < 2\alpha/m$ *for all* $k\in\mathbb{N}$;

(b) $f_i = \text{epi-lim}_{k\to+\infty} f_i^k$ *for each* $i=1,\cdots,m$;

(c) $\lim_{k\to+\infty}\varepsilon_i^k=0$ *for each* $i=0,1,\cdots,m$;

(d) *The sequence* $\{y^k\}_{k\geq 0}$ *is bounded.*

Then the sequence $\{x^k\}_{k\geq 0}$ *converges to the unique solution of* (P) *and any accumulation point of the sequence* $\{y^k\}_{k\geq 0}$ *is a solution of the Fenchel dual program* (D) *of* (P). □

Before giving the proof, let us notice that under the hypotheses of Theorem 2, the Fenchel dual program (D) of (P) admits a solution.

Proof. By hypothesis (b), for each $i=1,\cdots,m$, $\{f_i^k\}_{k\geq 0}$ is a sequence of closed proper convex functions epiconvergent to f_i. Then, by virtue of a theorem of Wijsman [29] (see, for example, Attouch [2, Thm. 3.18]), epiconvergence of their conjugate functions to that of f_i also occurs:

$$f_i^* = \underset{k\to+\infty}{\text{epi-lim}}\,(f_i^k)^* \quad \text{for each } i=1,\cdots,m.$$

But this is equivalent to (see (4.5))

$$g_i = \underset{k\to+\infty}{\text{epi-lim}}\, g_i^k \quad \text{for each } i=1,\cdots,m.$$

Using Theorem 6 of McLinden and Bergstrom [20], we obtain

$$G=\underset{k\to+\infty}{\text{epi-lim}}\, G^k, \tag{4.23}$$

where, of course,

$$G(y)=\sum_{i=1}^m g_i(y_i) \quad\text{and}\quad G^k(y)=\sum_{i=1}^m g_i^k(y_i) \quad \text{for all } y=(y_1,\cdots,y_m)\in\mathbb{R}^{m\cdot n}.$$

On the other hand, combining assumptions (c), (a), and the estimate (4.19) of Corollary 1 gives

$$\lim_{k\to+\infty}\|e^k\|=0. \tag{4.24}$$

Because f_0 is strongly convex with modulus α, it follows from (2.6) by a simple computation, that ∇f_0^* is Lipschitz continuous with constant $1/\alpha$, and consequently ∇F_0 is Lipschitz continuous with constant m/α. This fact, together with hypothesis (a), (4.18), (4.23), (4.24), and hypothesis (d) allow us to invoke Theorem 2 of Lemaire [18] to conclude that the Fenchel dual program (D) has at least one solution and that any accumulation point of $\{y^k\}_{k\geq 0}$ (its existence is guaranteed by hypothesis (d)) is a solution of (D). This proves the second part of the theorem.

To prove the first part of the theorem, let us begin by showing that the sequence $\{x^k\}_{k\geq 0}$ is bounded. To see this, note that the sequence $\{\sum_{i=1}^m y_i^k\}_{k\geq 0}$ is also bounded in $\mathbb{R}^n$ because $\|\sum_{i=1}^m y_i^k\|\leq\sqrt{m}\,\|y^k\|$ and $\{y^k\}_{k\geq 0}$ is bounded by assumption (d). Since ∇f_0^* is Lipschitz continuous with constant $1/\alpha$, we deduce that the sequence $\{\nabla f_0^*(\sum_{i=1}^m y_i^k)\}_{k\geq 0}$ is bounded. Hence, it follows from (4.6), (4.8) of Theorem 1, and hypothesis (c) that the sequence $\{x^k\}_{k\geq 0}$ is also bounded.

If it can be shown that any accumulation point of $\{x^k\}_{k\geq 0}$ solves (P), the first part of the theorem will be proven since (P) has a unique solution.

We therefore consider a subsequence $\{x^{k_j}\}_{j\geq 0}$ of $\{x^k\}_{k\geq 0}$ converging to $\bar{x}\in\mathbb{R}^n$. Since $\{y^k\}_{k\geq 0}$ is bounded, by taking a further subsequence, if necessary, we can find a point $\bar{y}=(\bar{y}_1,\cdots,\bar{y}_m)\in\mathbb{R}^{m\cdot n}$ such that $y^{k_j}=(y_1^{k_j},\cdots,y_m^{k_j})\rightarrow\bar{y}$. On the other hand,by (4.6) and (4.8) of Theorem 1, we have, for all $j\in\mathbb{N}$,

$$x^{k_j}=\nabla f_0^*(y_1^{k_j}+\cdots+y_m^{k_j})+e_0^{k_j} \tag{4.25}$$

with

$$\|e_0^{k_j}\|\leq\sqrt{2\varepsilon_0^{k_j}/\alpha}. \tag{4.26}$$

By passing to the limit $j\rightarrow+\infty$ in (4.25) and (4.26), we see that

$$\bar{x}=\nabla f_0^*\left(\sum_{i=1}^m \bar{y}_i\right) \tag{4.27}$$

because ∇f_0^* is continuous and $\varepsilon_0^{k_j}\rightarrow 0$ as $j\rightarrow+\infty$ (hypothesis (c)). It follows that

$$\sum_{i=1}^m \bar{y}_i=\nabla f_0(\bar{x}). \tag{4.28}$$

Now, by the first part of the proof, we know that $\bar{y}$ is a solution of (D). Therefore, we have the following characterization of optimality:

$$0\in\nabla F_0(\bar{y})+\partial G(\bar{y}).$$

This implies immediately that

$$0\in\nabla f_0^*\left(\sum_{i=1}^m \bar{y}_i\right)+\partial g_i(\bar{y}_i)\quad\text{for all } i=1,\cdots,m,$$

or equivalently (see (4.27)),

$$0\in\bar{x}-\partial f_i^*(-\bar{y}_i)\quad\text{for all } i=1,\cdots,m.$$

Thus (see [22, Thm. 23.5]),

$$-\bar{y}_i\in\partial f_i(\bar{x})\quad\text{for all } i=1,\cdots,m.$$

Hence we have

$$-\sum_{i=1}^m \bar{y}_i\in\sum_{i=1}^m \partial f_i(\bar{x})$$

so that (see (4.28))

$$0\in\nabla f_0(\bar{x})+\partial\left(\sum_{i=1}^m f_i\right)(\bar{x})$$

because $\sum_{i=1}^m \partial f_i(\bar{x})\subset\partial(\sum_{i=1}^m f_i)(\bar{x})$. This completes the proof of the theorem. □

In connection with the work of Han and Lou [13], it is worth mentioning that our Theorem 2 is a generalization and an improvement of Theorem 3.2 of [13]. It is a generalization because it suffices to take, in the theorem above: $\lambda_k=1/2a$ for all $k\in\mathbb{N}$, $f_i^k=f_i$ for all $i=1,\cdots,m$ and all $k\in\mathbb{N}$, and $\varepsilon_i^k=0$ for all $i=0,1,\cdots,m$ and all $k\in\mathbb{N}$, in order to obtain the convergence of the sequence $\{x^k\}_{k\in\mathbb{N}}$ to the optimal solution of (P). It is also an improvement because, not as in [13], the theorem above holds true without any regularity condition. Moreover, not as here, there are no convergence results in Theorem 3.2 of [13] concerning the sequence $\{y^k\}_{k\in\mathbb{N}}$.

Note that hypothesis (d) of Theorem 2 is crucial, and must be checked in each particular case. Nevertheless, we have the following convergence result.

THEOREM 3. *Under the same notation as in the preceding theorem, suppose the following*:

(a) $0<\underline{\lambda}\le\lambda_k\le\bar{\lambda}<2\alpha/m$ *for all* $k\in\mathbb{N}$;

(b) $f_i^k\le f_i^{k+1}$ *for each* $i=1,\cdots,m$ *and all* $k\in\mathbb{N}$, *and* $f_i=\sup_{k\in\mathbb{N}}f_i^k$ *for each* $i=1,\cdots,m$;

(c) $\lim_{k\to+\infty}\varepsilon_i^k=0$ *for each* $i=0,1,\cdots,m$;

(d) *the solution set of the Fenchel dual program* (D) *is a nonempty bounded set.*

Then any accumulation point of the sequence $\{y^k\}_{k\in\mathbb{N}}$ *is a solution of* (D) *and the sequence* $\{x^k\}_{k\in\mathbb{N}}$ *converges to the unique solution of* (P).

Proof. In view of Theorem 3.20 of Attouch [2], hypothesis (b) implies

$$f_i=\operatorname*{epi-lim}_{k\to+\infty}f_i^k\quad\text{for all } i=1,\cdots,m.$$

Consequently, as in the first part of the proof of Theorem 2, we have

$$G=\operatorname*{epi-lim}_{k\to+\infty}G^k.$$

On the other hand, as the sequence $\{f_i^k\}_{k\in\mathbb{N}}$ is monotone increasing, the sequence of their conjugates $\{(f_i^k)^*\}_{k\in\mathbb{N}}$ is monotone decreasing, so that the sequence $\{g_i^k\}_{k\in\mathbb{N}}$ is also monotone decreasing. Hence, from the definition of G^k, we have that

$$\{G^k\}_{k\in\mathbb{N}}\quad\text{is monotone decreasing.}\tag{4.29}$$

According to Theorem 3.20 of Attouch [2] again, we get

$$G=\operatorname{cl}\left(\inf_{k\in\mathbb{N}}G^k\right),\tag{4.30}$$

where cl denotes the closure (of the convex function $\inf_{k\in\mathbb{N}}G^k$).

Now, to show the theorem, it suffices to show that the sequence $\{y^k\}_{k\in\mathbb{N}}$ is bounded, because of Theorem 2. But first we deduce from assumption (d) that the problem (D) has a finite optimal value $\inf(\mathrm{D})\in\mathbb{R}$. Note that the solution set of (D) is also the level set

$$L=\{y\in\mathbb{R}^N\,|\,F_0(y)+G(y)\le\inf(D)\}.$$

As it is nonempty and bounded by hypothesis, it follows from Corollary 8.7.1 of [22] that the set

$$S_\gamma=\{y\in\mathbb{R}^N\,|\,F_0(y)+G(y)\le\gamma\}$$

is bounded for every $\gamma\in\mathbb{R}$. As a consequence, the function F_0+G is coercive; that is, we have

$$\lim_{\|y\|\to+\infty}(F_0(y)+G(y))=+\infty.\tag{4.31}$$

From Corollary 1, the fact that ∇F_0 is Lipschitz continuous and (4.29)-(4.31), we can conclude, by using Proposition 4(iii) of Lemaire [18], that the sequence $\{y^k\}_{k\in\mathbb{N}}$ is bounded, which completes the proof of the theorem. □

An example of functions f_i^k (for $i=1,\cdots,m$, and $k\in\mathbb{N}$) that satisfy assumption (b) of Theorem 3 has been given in § 2 (see (H1)-(H3)). Another instance is provided by the class of casting functions due to Wets [28, § 6].

In view of Theorem 3, it is important from a practical standpoint, to study when the solution set of the Fenchel dual program (D) is nonempty and bounded. To ensure this, we use the Fenchel duality theory. More precisely, we conclude this section on this matter with the following result.

THEOREM 4. *The notation being the same as above, if the condition*

$$\operatorname{int}(\operatorname{dom} f_1) \cap \cdots \cap \operatorname{int}(\operatorname{dom} f_m) \neq \emptyset \tag{4.32}$$

is satisfied, then the solution set of the Fenchel dual program (D) *of* (P) *is a nonempty bounded set.*

Before giving the proof, let us note that condition (4.32) holds true automatically if the functions f_i are finite real valued (for an example, see (1.2)).

Proof. By Theorem 27.1 of [22], we know that the solution set of the Fenchel dual program (D) of (P) is a nonempty bounded set if and only if

$$0 \in \operatorname{int}(\operatorname{dom}(F_0 + G)^*). \tag{4.33}$$

Therefore, to prove the theorem we need only to establish that hypothesis (4.32) implies that (4.33) holds true. Using Theorem 16.4 of [22], we have

$$\operatorname{dom}(F_0 + G)^* = \operatorname{dom} F_0^* + \operatorname{dom} G^* \tag{4.34}$$

because

$$\operatorname{ri}(\operatorname{dom} F_0) \cap \operatorname{ri}(\operatorname{dom} G) = \operatorname{ri}(\operatorname{dom} G) \neq \emptyset.$$

Start by determining the set dom F_0^*. To do this recall that

$$F_0(y) = F_0(y_1, \cdots, y_m) = f_0^*\left(\sum_{i=1}^m y_i\right).$$

So, using Theorem 16.3 of [22] and the fact that dom $f_0^* = \mathbb{R}^n$, we get

$$\operatorname{dom} F_0^* = \{(x_1, \cdots, x_m) \in \mathbb{R}^{m \cdot n} | x_1 = \cdots = x_m\}. \tag{4.35}$$

We now use hypothesis (4.32) to determine the set dom G^*. Let $\bar{x} \in \mathbb{R}^n$ be such that

$$\bar{x} \in \bigcap_{i=1}^m \operatorname{int}(\operatorname{dom} f_i).$$

This gives

$$0 \in -(\bar{x}, \cdots, \bar{x}) + \prod_{i=1}^m \operatorname{int}(\operatorname{dom} f_i). \tag{4.36}$$

Because $g_i(y_i) = f_i^*(-y_i)$ by definition, we have $g_i^*(x_i) = f_i(-x_i)$ for $i = 1, \cdots, m$. It now follows that dom $g_i^* = -\operatorname{dom} f_i$. Thus, from (4.36), we obtain

$$0 \in (\bar{x}, \cdots, \bar{x}) + \operatorname{int}\left(\prod_{i=1}^m \operatorname{dom} g_i^*\right).$$

Hence,

$$0 \in (\bar{x}, \cdots, \bar{x}) + \operatorname{int}(\operatorname{dom} G^*) \tag{4.37}$$

because

$$G^*(x_1, \cdots, x_m) = \sum_{i=1}^m g_i^*(x_i)$$

and

$$\operatorname{dom} G^* = \prod_{i=1}^{m} \operatorname{dom} g_i^*$$

(see McLinden [19, Lemma 4.3]).

On the other hand, the set $(\bar{x}, \cdots, \bar{x}) + \operatorname{int}(\operatorname{dom} G^*)$ is an open set. It follows from (4.37) and (4.35) that

$$0 \in (\bar{x}, \cdots, \bar{x}) + \operatorname{int}(\operatorname{dom} G^*) \subset \operatorname{dom} F_0^* + \operatorname{dom} G^*.$$

Hence, we deduce from (4.34) that

$$0 \in \operatorname{int}(\operatorname{dom}(F_0 + G)^*),$$

which is precisely (4.33) and this completes the proof. □

Note that condition (4.32) holds true automatically if the functions f_i are finite real valued (for an example, see (1.2)).

5. Conclusions. We have presented a convergent perturbed parallel decomposition method for minimizing the model objective function $f_0 + \sum_{i=1}^{m} f_i$. The method allows for approximate minimizations, duality gaps, and epigraphical perturbations during the iterative process.

At the present time, there are no rate-of-convergence results relating to our decomposition method. With respect to this topic the notion of variational semidistance due to Attouch and Wets [4] seems to be the main technical tool to employ. This is a subject for further research.

Acknowledgments. We thank Professors R. T. Rockafellar and R. Wets for helpful discussions. We are also indebted to Professor R. Wets for providing references [5], [26], and [28].

REFERENCES

[1] P. ALART AND B. LEMAIRE, *Penalization in non-classical convex programming via variational convergence*, Math. Programming, submitted.

[2] H. ATTOUCH, *Variational Convergence for Functions and Operators*, Pitman Research Notice in Mathematics, Pitman, London, 1984.

[3] H. ATTOUCH AND R. WETS, *Approximation and Convergence in Nonlinear Optimization*, Nonlinear Programming 4, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1981, pp. 367-394.

[4] ———, *Isometries for the Legendre-Fenchel transform*, Trans. Amer. Math. Soc., 296 (1986), pp. 33-60.

[5] ———, *Epigraphical analysis*, Tech. Report, University of California, Davis, CA, May 1988.

[6] A. AUSLENDER, *Numerical methods for non-differentiable convex optimization*, Math. Programming Stud., 30 (1987), pp. 102-126.

[7] A. AUSLENDER, J. P. CROUZEIX, AND P. FEDIT, *Penalty-proximal methods in convex programming*, J. Optim. Theory Appl., 55 (1987), pp. 1-21.

[8] J. F. BARTHELEMY, *Engineering applications of heuristic multilevel optimization methods*, NASA Tech. Memorandum 101504, NASA Langley Research Center, Hampton, VA, October 1988.

[9] H. BREZIS, *Opérateurs maximaux monotones et semi-groupes de contractions dans les espaces de Hilbert*, North-Holland, Amsterdam, 1973.

[10] G. COHEN, *Auxiliary problem principle and decomposition of optimization problems*, J. Optim. Theory Appl., 32 (1980), pp. 277-305.

[11] M. FUKUSHIMA, *A descent algorithm for nonsmooth convex programming*, Math. Programming, 30 (1984), pp. 163-175.

[12] S. P. HAN AND G. LOU, *A parallel algorithm for a class of convex programs*, SIAM J. Control Optim., 26 (1988), pp. 345-355.

[13] ———, *Some parallel decomposition algorithms for convex programming*, Tech. Report, Department of Mathematics, University of Illinois, Urbana, IL, September 1987.

[14] S. P. HAN, *A successive projection method*, Math. Programming, 40 (1988), pp. 1-14.

[15] K. KIWIEL, *Methods of Descent for Nondifferentiable Optimization*, Lecture Notes in Mathematics, Springer-Verlag, Berlin, 1985.

[16] L. S. LASDON, *Optimization Theorey for Large Systems*, MacMillan, New York, 1970.

[17] B. LEMAIRE, *The proximal algorithm*, presented at the Symposium on New Methods of Optimization and their Industrial Use, University of Pau-ELF Aquitaine, Pau, France, October 19-20, 1987.

[18] ———, *Coupling optimization methods and variational convergence*, in Trends in Mathematical Optimization, K. H. Hoffmann, J. B. Hiriart-Urruty, and J. Zowe, eds., International Series of Numerical Mathematics 84, Birkhäuser-Verlag, Basel, 1988, pp. 163-179.

[19] L. MCLINDEN, *Minimax problems, saddle functions and duality*, Ph.D. thesis, University of Washington, Seattle, WA, 1971.

[20] L. MCLINDEN AND R. C. BERGSTROM, *Preservation of convergence of convex sets and functions in finite dimensions*, Trans. Amer. Math. Soc., 268 (1981), pp. 127-142.

[21] H. MINE AND M. FUKUSHIMA, *Penalty function theory for general convex programming methods*, J. Optim. Theory Appl., 24 (1978), pp. 287-301.

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

[23] ———, *Monotone operators and the proximal point algorithm*, SIAM J. Control Optim., 14 (1976), pp. 877-898.

[24] R. T. ROCKAFELLAR AND R. WETS, *Variational systems, an introduction*, in Multifunctions and Integrands: Stochastic Analysis, Approximation and Optimization, G. Salinetti, ed., Lecture Notes in Mathematics, Vol. 1091, Springer-Verlag, Berlin, 1984, pp. 1-54.

[25] J.-J. STRODIOT AND V. H. NGUYEN, *On the numerical treatment of the inclusion* $0 \in \partial f(x)$, Topics in Nonsmooth Mechanics, J. J. Moreau, P. D. Panagiotopoulos, and G. Strang, eds., Birkhäuser-Verlag, Basel, Switzerland, 1988, pp. 267-294.

[26] P. VARAIYA AND R. WETS, *Stochastic dynamic optimization approaches and computation*, in Proc. 13th Internat. Symposium on Mathematical Programming, Tokyo, Japan, August 29-September 2, 1988.

[27] R. WETS, *Stochastic programming: solution techniques and approximation schemes*, in Mathematical Programming: The State of the Art, Bonn 1982, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1982, pp. 566-603.

[28] ———, *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., Wiley, Chichester, UK, 1980, pp. 376-403.

[29] R. A. WIJSMAN, *Convergence of sequences of convex sets, cones and functions*, II, Trans. Amer. Math. Soc., 123 (1966), pp. 32-45.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 848-865, July 1991

ON EIGENSTRUCTURE ASSIGNMENT BY GAIN OUTPUT FEEDBACK*

CALIXTE CHAMPETIER† AND JEAN-FRANCOIS MAGNI‡

Abstract. In this paper, the eigenstructure assignment of linear multivariable control systems is studied from a geometric point of view. For the class of systems in which the number of outputs plus the number of inputs exceeds the number of states, genericity properties relative to this problem are derived. It is shown, without any assumption on the genericity of the system, that the pole assignment can be carried out by choosing some closed-loop *eigenvectors* almost freely. The crucial point is that all the expected degrees of freedom in pole assignment are so described without redundancy, which fully justifies the practical interest of such techniques. Despite the technicality required for the derivation of intermediate results, the main result, which is an eigenstructure assignment algorithm, is very easy to implement because it is only based on the computation of certain sums and intersections of characteristic subspaces. Furthermore, it is shown how the basic tools developed here can be used to tackle the problem of finding a necessary and sufficient condition for exact pole assignment.

Key words. linear systems, multivariable control, output feedback, pole assignment, eigenvector assignment

AMS(MOS) subject classification. 93B

1. Introduction. Since the fundamental paper of Kimura [7], the pole assignment problem for systems satisfying $m+p>n$ (where n, m, p stand, respectively, for the numbers of states, inputs, and outputs), has received much investigation. It was shown in this paper that for a given triple satisfying the above condition, an *almost arbitrary* set of distinct closed-loop poles was assignable by gain output feedback. The use of the *characteristic subspaces*, which were introduced in this reference, enlightened various problems in linear system theory and was at the origin of the eigenvector assignment techniques as in Moore [13]. Simultaneously, a similar result appeared in Davison and Wang [2]. This result is less general because it is based on genericity assumptions with respect to the input and output matrices—therefore on the controllability and observability indices—of the considered system. In fact, as we will demonstrate in the present paper, the difficulties that arise for pole assignment are closely related to these indices (these difficulties are even more severe when it is not assumed that $m+p>n$). As the indices of practical systems do not often satisfy generic properties, this kind of assumption is to be avoided as far as possible. Among the various contributions that followed the fundamental work of Kimura, let us mention more specifically that of Fletcher [4] in which are discussed the proofs given in [7].

The simultaneous assignment of both eigenvalues and eigenvectors has also received much attention. In the specific case where the assumption $m+p>n$ is satisfied, Mielke and Liberty [12] proposed an algorithm based on the approach of [2] for freely assigning $p-1$ eigenvectors while the remaining degrees of freedom were used for assigning the $n-p+1$ other eigenvalues. It is worth noting that in this work as well as in the work of various other authors, only descriptions of eigenstructure assignment procedures are given, the cases in which the corresponding algorithms may fail are not studied.

* Received by the editors February 27, 1989; accepted for publication (in revised form) July 12, 1990. This research was supported by the Direction des Recherches, Etudes et Techniques of the French Délégation Générale à l'Armement.

† Matra-Espace, F31077, Toulouse Cedex, France. The research of this author was performed while he was with the Centre d'Etude et de Recherche de Toulouse.

‡ Centre d'Etude et de Recherche de Toulouse, BP 4025, F31055 Toulouse Cedex, France.

Another domain of interest which followed the contribution of Kimura is the problem of *exact* pole assignment. The first result obtained in this field is due to Fletcher [3], who proved that a set of n distinct real eigenvalues or a set of n distinct, nonreal, self-conjugate eigenvalues could exactly be assigned. The first result concerning the pathologies when a mixture of real and nonreal numbers are to be assigned is given in Magni [8] in terms of a necessary condition. It is shown in Magni [9] that this necessary condition is as well sufficient. This proof is given in the self-contained set of papers: [5], [6], [10].

In this paper we first study, from a geometric point of view, the properties of the characteristic subspaces. Some of the results obtained are also useful for more general pole assignment problems as, for instance, those presented in Magni and Champetier [11]. Our main result concerns the genericity relative to the choice of the eigenvalues within their corresponding characteristic subspaces. We prove that the eigenvalues/eigenvectors assignment can be performed iteratively, almost any choice at each step of the algorithm being such that the algorithm will not fail until $p-1$ eigenvectors and n eigenvalues are assigned. We show as well that the set of pathologic eigenvalues is finite and the cardinal of this set is estimated from above. A numerical example shows that this set may not be empty. It is in this study of the genericity of the choice of the degrees of freedom exhibited by the proposed algorithm that our contribution with respect to previous results of the literature lies. An advantage of the approach adopted here is that, despite some difficulties arising in some proofs, the final assignment procedure is extremely simple to implement and only the computation of certain intersections and sums of characteristic subspaces is required. The intermediate result concerning the number of pathological eigenvalues which might appear during the assignment of the $p-1$ eigenvectors allows us to provide an alternative proof of the above-mentioned necessary and sufficient condition for exact pole assignment.

2. Characteristic subspaces. In this section, the definitions and various properties of the (A, B) and (C, A)-characteristic subspaces are recalled. We shall consider a complete (i.e., controllable and observable) triple (A, B, C), where the matrices B and C are assumed to be of full rank:

$$\mathcal{U} \xrightarrow{B} \mathcal{X} \xrightarrow{A} \mathcal{X} \xrightarrow{C} \mathcal{Y}.$$

First, let us recall the definition of these subspaces.

DEFINITION 2.1. For any complex number λ, the *(A, B)-characteristic subspaces* $\mathcal{S}(\lambda)$ is defined by $\mathcal{S}(\lambda) \triangleq (A-\lambda I)^{-1} \operatorname{Im} B$ and the *(C, A)-characteristic subspaces* $\mathcal{T}(\lambda)$ by $\mathcal{T}(\lambda) \triangleq (A-\lambda I) \operatorname{Ker} C$.

(Note that with "$(A-\lambda I)^{-1}$" being applied to a vector space, it is not necessary that the matrix $(A-\lambda I)$ be invertible to define $\mathcal{S}(\lambda)$ as above.) For notational conveniences, we shall denote $\bar{\mathbb{C}}$ the following set $\{\mathbb{C} \cup \{\infty\}\}$. $\mathcal{S}(\infty)$ will stand for $\operatorname{Im} \beta$ and $\mathcal{T}(\infty)$ for $\operatorname{Ker} C$. Note that for any vector $s \in \mathcal{S}(\lambda)$ it is clear that there exists a unique vector w called the *input direction corresponding to s* defined by $(A-\lambda I)s = -Bw$. In this paper, the *input directions* will be used only for explicit computations of gain feedback matrices.

Elementary properties of the characteristic subspaces will be needed in this paper. Their proofs (scattered in various papers, see, for instance, [3], [8], [11], [18]) can be derived in a straightforward way by considering the Brunovski canonical form. Let us state some of them. If $m<n$, $\mathcal{S}(\lambda) \neq \mathcal{S}(\lambda')$ for any $(\lambda, \lambda') \in \bar{\mathbb{C}}^2$ such that $\lambda \neq \lambda'$. The dimension of $\mathcal{S}(\lambda) \cap \mathcal{S}(\lambda')$ is equal to the number or controllability indices equal to one and the dimension of $\mathcal{S}(\lambda)+\mathcal{S}(\lambda')$ follows easily. If $p<n$, $\mathcal{T}(\lambda) \neq \mathcal{T}(\lambda')$ for any

$(\lambda, \lambda') \in \bar{\mathbb{C}}^2$ such that $\lambda \neq \lambda'$. Within some proofs given in this paper (Lemma 3.5, Lemma 4.3 third case), the knowledge of the characteristic subspaces on the Brunovski canonical form will be required and the reader will be referred to the literature cited above.

Both subspaces defined below will play a key role in the sequel:

$$(1)\qquad \mathscr{B}_0 = \bigcap_{\lambda \in \bar{\mathbb{C}}} \mathscr{S}(\lambda), \qquad \mathscr{C}_0 = \sum_{\lambda \in \bar{\mathbb{C}}} \mathscr{T}(\lambda).$$

Let p_0 and m_0 denote, respectively, the numbers of observability and controllability indices that are equal to one. The following properties of these subspaces are well known.

LEMMA 2.2.

$$\dim \mathscr{B}_0 = m_0 \textit{ and } \mathscr{B}_0 = \mathscr{S}(\lambda) \cap \mathscr{S}(\lambda') \textit{ for any } (\lambda, \lambda') \in \bar{\mathbb{C}}^2 \textit{ s.t. } \lambda \neq \lambda';$$

$$\dim \mathscr{C}_0 = n - p_0 \textit{ and } \mathscr{C}_0 = \mathscr{T}(\lambda) + \mathscr{T}(\lambda') \textit{ for any } (\lambda, \lambda') \in \bar{\mathbb{C}}^2 \textit{ s.t. } \lambda \neq \lambda'.$$

Note that for $\lambda = \infty$ and $\lambda' = 0$, we have $\mathscr{B}_0 = \operatorname{Im} B \cap A^{-1} \operatorname{Im} B$ and $\mathscr{C}_0 = \operatorname{Ker} C + A \operatorname{Ker} C$. Finally, we recall two lemmas, the proofs of which can be found in [8] or [10].

LEMMA 2.3. *If $n = m + p - 1$ with $m < n$ and $p < n$, then, there exists no more than one real number $\tilde{\lambda}$ such that*

$$\mathscr{C}_0 \subset \mathscr{S}(\tilde{\lambda}).$$

LEMMA 2.4. *If $p_0 = p - 1$, then $\mathscr{C}_0 = \bigcup_{\lambda \in \bar{\mathbb{C}}} \mathscr{T}(\lambda)$.*

The vocabulary and the classical notations of the geometric approach will be extensively used. The reader is referred to the basic literature in this field [17], [16]. Let us first recall the definition of the *induced* and *restricted* subsystems. If $\mathscr{S}$ is (A, B)-invariant and (C, A)-invariant, we can define the subsystem *induced modulo* $\mathscr{S}$: $(A(\operatorname{mod} \mathscr{S}), B(\operatorname{mod} \mathscr{S}), C(\operatorname{mod} \mathscr{S}))$ and the subsystem *restricted to* $\mathscr{S}$: $(A|\mathscr{S}, B|\mathscr{S}, C|\mathscr{S})$ by the commutative diagram of Fig. 2.1. Following [1] and [14] we can state the following property. Let (A, B, C) be a complete triple and $\mathscr{S}$ an (A, B)-invariant and (C, A)-invariant subspace. Then, there exists at least one output feedback K which satisfies $(A + BKC)\mathscr{S} \subset \mathscr{S}$ (such a feedback will be said to be *admissible*).

$$\begin{tikzcd}
B^{-1}\mathcal{S} \arrow[r, "B|\mathcal{S}"] \arrow[d, "\iota'"] & \mathcal{S} \arrow[r, "A|\mathcal{S}"] \arrow[d, "\iota"] & \mathcal{S} \arrow[d, "\iota"] \arrow[dr, "C|\mathcal{S}"] & \\
\mathcal{U} \arrow[r, "B"] \arrow[dr, "B(\mathrm{mod}\,\mathcal{S})"'] & \mathcal{X} \arrow[r, "A"] \arrow[d, "\pi"] & \mathcal{X} \arrow[r, "C"] \arrow[d, "\pi"] & \mathcal{Y} \arrow[d, "\pi'"] \\
 & \mathcal{X}/\mathcal{S} \arrow[r, "A(\mathrm{mod}\,\mathcal{S})"] & \mathcal{X}/\mathcal{S} \arrow[r, "C(\mathrm{mod}\,\mathcal{S})"] & \mathcal{Y}/C\mathcal{S}
\end{tikzcd}$$

FIG. 2.1. *Definition of the induced and restricted subsystems.*

Moreover, for any admissible feedback K, we have $\Lambda_F \cup \Lambda'_F \subset \sigma(A+BKC)$ where Λ_F and Λ'_F are the fixed spectra defined by the following lattice:

$$(2) \qquad 0 \longrightarrow \mathcal{R}^*_{\mathcal{S}} \xrightarrow{\Lambda_F} \mathcal{S} \xrightarrow{\Lambda'_F} \mathcal{N}^*_{\mathcal{S}} \longrightarrow \mathcal{X}.$$

Furthermore, Λ_F is the set of the invariant zeros of the triple (A, S, C) where S is any mapping such that $\operatorname{Im} S = \mathcal{S}$ and Λ'_F is the set of the invariant zeros of the triple (A, B, E) where E is any mapping such that $\operatorname{Ker} E = \mathcal{S}$. In fact, it is a more specific version of this result which will be used.

PROPOSITION 2.5. *Let $\Lambda = \{\lambda_1, \cdots, \lambda_n\}$ be a set of complex numbers. Assume that there exists a subspace $\mathcal{S}$ satisfying the following properties*:

(i) $\dim \mathcal{S} = p$;

(ii) $\mathcal{S} = \operatorname{Span}\{s_1, \cdots, s_p\}$ *with* $s_i \in \mathcal{S}(\lambda_i)$ *for* $i = 1, \cdots, p$;

(iii) $\mathcal{N}^*_{\mathcal{S}} = \mathcal{X}$ *with* $\Lambda'_F = \{\lambda_{p+1}, \cdots, \lambda_n\}$ (*cf. lattice* (2)).

Then, there exists a unique output feedback gain K satisfying

$$\sigma(A+BKC) = \Lambda \quad \textit{and} \quad (A+BKC)s_i = \lambda_i s_i \quad \textit{for } i = 1, \cdots, p.$$

This gain is given by $K = [w_1, \cdots, w_p](C[s_1, \cdots, s_p])^{-1}$ where the vectors w_i are the input directions corresponding to the vectors s_i.

Proof. It is straightforward to show that $\mathcal{N}^*_{\mathcal{S}} = \mathcal{X}$ is equivalent to $S + \operatorname{Ker} C = \mathcal{X}$. Since $\dim \mathcal{S} = p$, we have $\mathcal{S} \cap \operatorname{Ker} C = 0$. So, $\mathcal{S}$ is (C, A)-invariant. Consider the input directions w_i corresponding to s_i. As $\mathcal{S} \cap \operatorname{Ker} C = 0$, there exists a unique gain K such that $[w_1, \cdots, w_p] = KC[s_1, \cdots, s_p]$ which is equivalent to $(A+BKC)s_i = \lambda_i s_i$ for $i = 1, \cdots, p$. Therefore, this gain is admissible, and $\Lambda'_F \subset \sigma(A+BKC)$. Hence, $\sigma(A+BKC) = \Lambda$. □

The construction of the subspace $\mathcal{S}$ given in Proposition 2.5 can be used in order to assign the whole spectrum. But the third condition on $\mathcal{S}$ remains to be clarified.

THEOREM 2.6. *Let (C, A) be an observable pair, $\mathcal{S}$ a (C, A)-invariant subspace defining the lattice*

$$0 \longrightarrow \mathcal{S} \xrightarrow{\Lambda'_F} \mathcal{N}^*_{\mathcal{S}} \longrightarrow \mathcal{X}.$$

The following properties are equivalent:

(i) $\beta \in \Lambda'_F$;

(ii) $\dim \mathcal{S} \cap \mathcal{T}(\beta) \neq \dim \mathcal{S} \cap \operatorname{Ker} C$;

(iii) $\dim \mathcal{S} \cap \mathcal{T}(\beta) > \dim \mathcal{S} \cap \operatorname{Ker} C$.

More precisely, if $\beta \in \Lambda'_F$, then $\dim \mathcal{S} \cap \mathcal{T}(\beta) = \dim \mathcal{S} \cap \operatorname{Ker} C + r(\beta)$, where $r(\beta)$ is the geometrical multiplicity of the eigenvalue β.

Proof. Let $P(\beta)$ be the polynomial matrix defined by

$$P(\beta) = \begin{bmatrix} A - \beta I & S \\ C & 0 \end{bmatrix}.$$

Let us consider, for a given complex number β, the mapping F_β defined by

$$F_\beta: \quad \mathcal{S} \cap \mathcal{T}(\beta) \longrightarrow \operatorname{Ker} P(\beta),$$

$$x \longrightarrow \begin{bmatrix} v \\ w \end{bmatrix}$$

where $v \in \operatorname{Ker} C$, $x = (A - \beta I)v$, $x = -Sw$. As $x \in \mathcal{S}$ and $\mathcal{S}$ is of full rank, w exists and is unique; as $x \in \mathcal{T}(\beta)$, there exists a vector $v \in \operatorname{Ker} C$ such that $x = (A - \beta I)v$. This vector is unique since the pair (A, C) is observable. Then the mapping F_β is well defined. Furthermore, it is clear that $F_\beta(x) \in \operatorname{Ker} P(\beta)$. It is straightforward to show that F_β is linear and bijective. So,

$$\text{for all } \beta \in \mathbb{C}, \quad \dim \mathcal{S} \cap \mathcal{T}(\beta) = \dim \operatorname{Ker} P(\beta).$$

The rank property of $P(\beta)$ allows us to state that, if β_0 is an invariant zero of the triple (A, S, C), for all β which is not such a zero we have

$$\dim \mathcal{S} \cap \mathcal{T}(\beta_0) = \dim \mathcal{S} \cap \mathcal{T}(\beta) + r(\beta_0)$$

where $r(\beta_0)$ is the geometric multiplicity of β_0. As $r(\beta_0) > 0 \Leftrightarrow \beta_0 \in \Lambda'_F$, it suffices to show that $\dim \mathcal{S} \cap \operatorname{Ker} C = \dim \mathcal{S} \cap \mathcal{T}(\beta)$ to prove the proposition. $\mathcal{S}$ being (C, A)-invariant, there exists an input injection G such that $(A + GC)\mathcal{S} \subset \mathcal{S}$. For any complex number $\beta \notin \sigma(A + GC)(\Rightarrow \notin \Lambda'_F)$, $(A + GC - \beta I)$ is bijective and $(A + GC - \beta I)\mathcal{S} = \mathcal{S}$. Furthermore, the (C, A)-characteristic subspaces being invariant by input injection, $(A + GC - \beta I)\mathcal{T}(\beta) = \mathcal{T}(\beta)$. It follows that $(A + GC - \beta I)(\mathcal{S} \cap \operatorname{Ker} C) = \mathcal{S} \cap \mathcal{T}(\beta)$. □

Comment 1. From Theorem 2.6 and its dual counter part, it is easy to *characterize controllability and complementary observability subspaces in terms of characteristic subspaces.* Clearly, if (C, A) is an observable pair and $\mathcal{S}$ is (C, A)-invariant, $\mathcal{S}$ is a complementary observability subspace if and only if for all $\beta \in \mathbb{C}$, $\dim \mathcal{S} \cap \mathcal{T}(\beta) = \dim \mathcal{S} \cap \operatorname{Ker} C$. If (A, B) is a controllable pair and $\mathcal{S}$ is (A, B)-invariant, $\mathcal{S}$ is a controllability subspace if and only if for all $\lambda \in \mathbb{C}$, $\dim \mathcal{S} \cap \mathcal{S}(\lambda) = \dim \mathcal{S} \cap \operatorname{Im} B$. Note that this result can be viewed as a corollary of Theorem 7.1 of [15].

Finally, we will need the following results.

LEMMA 2.7. *Let $\mathcal{S}$ be (C, A) and (A, B)-invariant, $\tilde{\mathcal{S}}(\lambda)$ and $\tilde{\mathcal{T}}(\beta)$ denote the characteristic subspaces of $(A + BKC, B, C)$ induced modulo $\mathcal{S}$ (where K is an admissible feedback), and Λ_F denote the fixed spectrum corresponding to the lattice* (2).

(i) $\operatorname{Im}(B(\operatorname{mod} \mathcal{S})) = \pi(\operatorname{Im} B)$;

(ii) $\operatorname{Ker}(C(\operatorname{mod} \mathcal{S})) = \pi(\operatorname{Ker} C)$;

(iii) $\tilde{\mathcal{S}}(\lambda) = \pi(\mathcal{S}(\lambda))$ *for all* $\lambda \in \mathbb{C} \backslash \Lambda_F$;

(iv) $\tilde{\mathcal{T}}(\beta) = \pi(\mathcal{T}(\beta))$ *for all* $\beta \in \mathbb{C}$.

Proof. Let us denote $\tilde{B} = B(\operatorname{mod} \mathcal{S})$ and $\tilde{C} = C(\operatorname{mod} \mathcal{S})$. $\tilde{B} = \pi B$ yields (i). From the commutativity of the diagram of Fig. 2.1, we have

$$\operatorname{Ker} \tilde{C} = \pi \operatorname{Ker}(\tilde{C}\pi) = \pi \operatorname{Ker}(\pi' C) = \pi(\mathcal{S} + \operatorname{Ker} C) = \pi(\operatorname{Ker} C),$$

hence (ii) is satisfied. Moreover, it is easy to see that $\pi(\mathcal{S}(\lambda)) \subset \tilde{\mathcal{S}}(\lambda)$. In view of the dual counterpart of Theorem 2.6 from the relation

$$\dim \pi(\mathcal{S}(\lambda)) = \dim \mathcal{S}(\lambda) - \dim \mathcal{S} \cap \mathcal{S}(\lambda),$$

we have

$$\dim \pi(\mathcal{S}(\lambda)) = \dim \operatorname{Im} B - \dim \mathcal{S} \cap \operatorname{Im} B,$$

that is,

$$\dim \pi(\mathcal{S}(\lambda)) = \dim \pi(\operatorname{Im} B) = \dim \operatorname{Im} \tilde{B} = \dim \tilde{\mathcal{S}}(\lambda).$$

Part (iv) can be proved in the same manner by using (ii). □

3. Degrees of freedom in pole assignment by output feedback. Our main result concerning the available freedom for eigenvalue/eigenvector assignment is stated in Theorem 3.1 below. The whole section is devoted to its proof. To describe our results, a notion of iterative genericity is required. A property *depending on an iterative choice* in linear spaces will be said *iteratively generic* if it is generic at the first step in the sense where it is true except perhaps on a *thin* subset (i.e., a semialgebraic subset of nonmaximal dimension) of the first linear subspace, then, a nonpathologic choice being performed, the property remains generic with respect to the choice of a second vector in the second space, and so on.

3.1. Statement of the main result and description of the algorithm. The eigenstructure assignment procedure presented here requires a special treatment for the assignment of the pth eigenvector. This is considerably simplified if it is assumed that this special eigenvector corresponds to a real eigenvalue. Otherwise polynomial approaches must be used. So, it will be assumed that the set of eigenvalues to be assigned satisfies a slightly restrictive condition of admissibility. A finite set $\Lambda = \{\lambda_1, \cdots, \lambda_n\} \subset \mathbb{C}$ is said to be *admissible* if $\lambda_i \neq \lambda_j$ for $i \neq j$ and if there exists a partition of Λ consisting of three self-conjugate disjoint subsets of the form $\Lambda_1 = \{\lambda_1, \cdots, \lambda_{p-1}\}$, $\{\lambda_p\}$ and $\Lambda_2 = \{\lambda_{p+1}, \cdots, \lambda_n\}$. Now we can state our main result.

THEOREM 3.1. *Let (A, B, C) be a complete triple in $\mathbb{R}^{n\times n} \times \mathbb{R}^{n\times m} \times \mathbb{R}^{p\times n}$ satisfying $m+p>n$ and Λ be an admissible set of n complex numbers. Then (except perhaps for a finite subset of values of $\lambda_1, \cdots, \lambda_{p-1}$) the property*

$$\dim \mathcal{S}'(\lambda_p) = m+p-n$$

is iteratively generic with respect to the selection of vectors $s_i \in \mathcal{S}(\lambda_i)$, $i = 1, \cdots, p-1$ such that $s_i = \overline{s_j}$ if $\lambda_i = \overline{\lambda_j}$, where $\mathcal{S}'(\lambda_p) \triangleq \mathcal{S}(\lambda_p) \cap (\mathcal{T}(\lambda_{p+1}) + \mathcal{S}_{p-1}) \cap \cdots \cap (\mathcal{T}(\lambda_n) + \mathcal{S}_{p-1})$ with $\mathcal{S}_{p-1} \triangleq \langle s_1, \cdots, s_{p-1}\rangle$. Furthermore, given a generic vector s_p in $\mathcal{S}'(\lambda_p)$, we have $\langle s_1, \cdots, s_p\rangle \cap \operatorname{Ker} C \neq 0$ and the output feedback K given by

$$K = [w_1 \cdots w_p](C[s_1, \cdots, s_p])^{-1}$$

(where w_i is the input direction associated to s_i) is the unique real gain such that $\sigma(A+BKC) = \Lambda$ with $s_1, \cdots, s_p$ as eigenvectors associated to $\lambda_1, \cdots, \lambda_p$.

For synthesizing modal control laws, it is very important to have at one's disposal, without redundancy, all the degrees of freedom remaining after having chosen the pole to be assigned. Note that various pole assignment methods given in the literature waste degrees of freedom in "linearizing" the problem. Here the problem considered is naturally "linear." Let us check that all the degrees of freedom appearing in Theorem 3.1 are available without redundancy. There are mp entries in the gain matrix. Therefore after pole assignment it must remain exactly $mp-n$ free parameters. The choice of the $p-1$ first eigenvectors performed in m-dimensional subspaces, uses $(p-1)(m-1)$ degrees of freedom. Furthermore, the choice of the pth eigenvector being performed in an $(m+p-n)$-dimensional subspace, $m+p-n-1$ additional degrees of freedom are used. It is easy to check that $mp-n = (p-1)(m-1)+(m+p-n-1)$.

The eigenvectors s_i appearing in the above theorem are recursively chosen in order to construct the lattice given in Fig. 3.1 (in this lattice, $\mathcal{S}_i$ denotes the subspace Span $\{s_1, \cdots, s_i\}$; if λ_i is complex, we take $\lambda_{i+1} = \overline{\lambda_i}$ and $s_{i+1} = \overline{s_i}$). From Proposition 2.5, it is clear'that Theorem 3.1 will be proved if we can construct a subspace $\mathcal{S}_p$ satisfying the lattice of Fig. 3.1 for a generic choice of the vectors $s_1, \cdots, s_{p-1}$. The proof of such a result is the object of the sequel of this section. The lattice diagram (2) suggests that when $\mathcal{S}_i$ is made invariant, the set of assigned poles is exactly "$\Lambda_F \cup \Lambda'_F$." Note that in Fig. 3.1 the maximal controllability subspaces are not specified. It will appear in the proof given below (see (5)) that for $i = 1, \cdots, n-p$ we have $\mathcal{R}^*_{\mathcal{S}_i} = 0$. For $i > n-p$, $\mathcal{R}^*_{\mathcal{S}_i}$ is no longer equal to zero. But in view of Proposition 2.5, the spectrum of the subsystem restricted to $\mathcal{S}_i$ is nevertheless freely assignable by choosing a *special* gain feedback instead of considering *any* admissible gain feedback.

$$\begin{tikzcd}
0 \arrow[r, "\{\lambda_1\}"] & \mathcal{S}_1 = \mathcal{N}^*_{\mathcal{S}_1} \arrow[r] \arrow[d] & \mathcal{X} \\
0 \arrow[r, "\{\lambda_1, \lambda_2\}"] & \mathcal{S}_2 = \mathcal{N}^*_{\mathcal{S}_2} \arrow[r] \arrow[d] & \mathcal{X} \\
0 \arrow[r, "\{\lambda_1, \cdots, \lambda_{p-1}\}"] & \mathcal{S}_{p-1} = \mathcal{N}^*_{\mathcal{S}_{p-1}} \arrow[r] \arrow[d] & \mathcal{X} \\
0 \arrow[r, "\{\lambda_1, \cdots, \lambda_p\}"] & \mathcal{S}_p \arrow[r, "\{\lambda_{p+1}, \cdots, \lambda_n\}"] & \mathcal{X} = \mathcal{N}^*_{\mathcal{S}_p}
\end{tikzcd}$$

FIG. 3.1. *Illustration of the algorithm by a lattice diagram.*

3.2. Initialization of the construction of the lattice. The subspaces $\mathcal{S}_i$, $i = 1, \cdots, p-1$ appearing in Fig. 3.1 will be iteratively constructed. The main difficulty lies in the initialization of the procedure. The construction of $\mathcal{S}_{i+1}$ from $\mathcal{S}_i$ is easier and will be treated later. The following proposition deals with the initialization of the procedure in the case where the involved eigenvalue is real.

PROPOSITION 3.2. *Let (A, B, C) be a complete triple satisfying $m+p>n$. There exists no more than one real number $\tilde{\lambda}$ such that for any $\lambda \in \mathbb{R} \backslash \{\tilde{\lambda}\}$, for almost every vector $s \in \mathcal{S}(\lambda)$, the subspace $\mathcal{S} = \langle s \rangle$ satisfies*

(3) $$\mathcal{S} \cap \operatorname{Ker} C = 0,$$

(4) $$\textit{for all } \beta \in \mathbb{C}, \quad \mathcal{S} \cap \mathcal{T}(\beta) = 0,$$

(5) $$\textit{if } m < n, \quad \mathcal{S} \cap \operatorname{Im} B = 0.$$

Proof. Clearly, if $m<n$, $\mathcal{S}(\lambda) \neq \operatorname{Im} B$ for all $\lambda \in \mathbb{C}$. So, $\mathcal{S}(\lambda) \cap \operatorname{Im} B$ is a proper subspace of $\mathcal{S}(\lambda)$. In view of the dimensions of $\mathcal{S}(\lambda)$ and $\operatorname{Ker} C$, it is clear that $\mathcal{S}(\lambda) \cap \operatorname{Ker} C$ is a proper subspace of $\mathcal{S}(\lambda)$. The proposition will be demonstrated if we show that the set

$$\mathcal{S}(\lambda) \cap \bigcup_{\beta \in \mathbb{C}} \mathcal{T}(\beta)$$

is thin in $\mathcal{S}(\lambda)$. Let p_0 be the number of observability indices equal to one. Two cases must be considered.

1) $p - p_0 = 1$. Then, from Lemma 2.4

$$\operatorname{Ker} C \cup \bigcup_{\beta \in \mathbb{C}} \mathcal{T}(\beta) = \mathcal{C}_0.$$

It remains to show that $\mathcal{C}_0 \cap \mathcal{S}(\lambda)$ is a proper subspace of $\mathcal{S}$. Assume that $\mathcal{S}(\lambda) \subset \mathcal{C}_0$. As $n \leqq m+p-1$, we have $n - p_0 \leqq m$, i.e., dim $\mathcal{C}_0 \leqq$ dim $\mathcal{S}(\lambda)$. Hence,

(6) $$\mathcal{C}_0 = \mathcal{S}(\lambda).$$

So necessarily $n = m+p-1$. From Lemma 2.3, there exists no more than one value $\tilde{\lambda}$ of λ such that (6) is satisfied.

2) $p-p_0>1$. From now on, λ is assumed to be different from $\tilde{\lambda}$. From the definition of the (C, A)-characteristic subspaces, there exist p polynomials P_i, $i=1, \cdots, p$ in $n+1$ complex variables such that

$$(7) \qquad x=(x_1, \cdots, x_n)^T \in \mathcal{T}(\beta) \quad \text{iff } P_i(x_1, \cdots, x_n, \beta)=0 \quad \text{for } i=1, \cdots, p.$$

Consider the set $\mathcal{S}(\lambda, \beta) \subset \mathbb{C}^{n+1}$ consisting of the points $(x_1, \cdots, x_n, \beta)$ satisfying

$$x=(x_1, \cdots, x_n)^T \in \mathcal{S}(\lambda) \cap \mathcal{T}(\beta).$$

This set is a closed algebraic subset in $\mathbb{C}^{n+1}$ defined by the $n-m+p$ polynomial equations: those appearing in (7) together with the linear equations in $x_1, \cdots, x_n$ satisfied by any vector x belonging to $\mathcal{S}(\lambda)$. Let r be the algebraic dimension of this subset and assume that $r>m-1$. This means that there exists no more than $n+1-m$ algebraically independent polynomials P_i', $i=1, \cdots, n+1-m$ such that

$$(x, \beta) \in \mathcal{S}(\lambda, \beta) \quad \text{iff } P_i'(x, \beta)=0.$$

Then, for any $\beta_0 \in \mathbb{C}$, the subspace $\mathcal{S}(\lambda) \cap \mathcal{T}(\beta_0)$ is of dimension $\geqq m-1$ (this subspace is described by the equations $\beta=\beta_0$, $P_i'(x, \beta_0)=0$ for $i=1, \cdots, n-m+1$). As $\dim \mathcal{T}(\beta_0)=n-p<m=\dim \mathcal{S}(\lambda)$, we must have $\mathcal{T}(\beta_0) \subset \mathcal{S}(\lambda)$ for all β_0, i.e. from Lemma 2.2, $\mathcal{C}_0 \subset \mathcal{S}(\lambda)$. From Lemma 2.3, $\lambda=\tilde{\lambda}$ which contradicts the hypothesis. So $\dim \mathcal{S}(\lambda, \beta) \leqq m-1$, and the set $\mathcal{S}(\lambda) \cap \bigcup_{\beta \in \mathbb{C}} \mathcal{T}(\beta)$, which is the projection of $\mathcal{S}(\lambda, \beta)$ on $\mathcal{S}(\lambda)$, is contained in a semialgebraic subset of dimension $\leqq m-1$. □

The complex case is much more intricate.

PROPOSITION 3.3. *Let (A, B, C) be a complete triple satisfying $m+p>n$. There exists a finite subset $\Xi \subset \mathbb{C}$ such that for all $\lambda \in \mathbb{C} \backslash (\mathbb{R} \cup \Xi)$, for almost every vector $s \in \mathcal{S}(\lambda)$, the subspace $\mathcal{S}=\langle s, \bar{s} \rangle$ satisfies*

$$(8) \qquad \dim \mathcal{S}=2,$$

$$(9) \qquad \mathcal{S} \cap \operatorname{Ker} C=0,$$

$$(10) \qquad \mathcal{S} \cap \mathcal{T}(\beta)=0 \quad \textit{for all } \beta \in \mathbb{C},$$

$$(11) \qquad \textit{if } m<n-1, \quad \mathcal{S} \cap \operatorname{Im} B=0,$$

$$(12) \qquad \textit{if } m \geqq n-1, \quad \mathcal{S}+\operatorname{Im} B=\mathcal{X}.$$

The cardinal of Ξ can be estimated from above as follows:

$$\textit{if } p \geqq 4, \quad \operatorname{Card} \Xi \leqq 2(n-p),$$

$$\textit{if } p=3, \quad \operatorname{Card} \Xi \leqq 4(n-p).$$

Proof. In a first step, a vector in $\mathcal{S}(\lambda)$ such that (10) is satisfied will be constructed (such a vector will be called *admissible*) and an estimation of Card Ξ will be given. In a second step, we will show that almost every vector in $\mathcal{S}(\lambda)$ is admissible.

Step 1. Construction of an admissible vector and estimation of Card Ξ. The basic idea for constructing an admissible vector consists in decreasing the order of the system to deal with. This artifice is needed to avoid the treatment of numerous special cases. In fact, in that way, only three cases (see Lemma 3.5) are to be considered. To decrease the order, we are going to use a result (Proposition 3.10) which will be proved in the next section. This proposition is used here in a case where only real eigenvalues are assigned; in this case its proof is independent of the results stated in Proposition 3.3 (it depends only on Proposition 3.2). As $n-p \leqq m-1$, from Proposition 3.2 and 3.10, $n-p$ distinct *real* eigenvalues of the triple (A^T, C^T, B^T) can be assigned by output feedback. More precisely, this pole assignment can be achieved by choosing a subspace $\mathcal{Q}$ satisfying

- $\dim \mathcal{Q} = n-p$, $\mathcal{Q}=\bar{\mathcal{Q}}$;
- $\mathcal{Q}$ is a complementary observability subspace of the pair (A^T, B^T);
- $\mathcal{Q} \cap \operatorname{Im} C^T = 0$;
- $\mathcal{Q} \cap \operatorname{Ker} B^T = 0$;
- There exists a feedback K such that the spectrum of $(A^T + C^T K^T B^T)|\mathcal{Q}$ is a set, denoted by Λ'_F, of $n-p$ distinct real numbers arbitrarily chosen.

With respect to the triple (A, B, C), the subspace $\mathcal{S}=\mathcal{Q}^{\perp}$ satisfies

- $\dim \mathcal{S} = p$, $\mathcal{S}=\bar{\mathcal{S}}$;
- $\mathcal{S}$ is a controllability subspace;
- $\mathcal{S} + \operatorname{Ker} C = \mathbb{C}^n$ (hence $\mathcal{S} \cap \operatorname{Ker} C = 0$);
- $\mathcal{S} + \operatorname{Im} B = \mathbb{C}^n$ (hence $\dim \mathcal{S} \cap \operatorname{Im} B = m+p-n$);
- $\sigma(A+BKC)(\operatorname{mod} \mathcal{S}) = \Lambda'_F$;
- Λ'_F corresponds to the fixed spectrum of the lattice (2) defined by $\mathcal{S}$.

From now on, the spectrum Λ'_F and the gain K will be fixed. Consider the triple $(\hat{A}, \hat{B}, \hat{C})$, restriction of the triple $(A+BKC, B, C)$ to $\mathcal{S}$. The number of states, outputs, inputs of this system are, respectively, $n'=p$, $p'=p$ (because $\mathcal{S} \cap \operatorname{Ker} C = 0$), $m' = m+p-n$. Obviously, the characteristic subspaces of the triples $(A+BKC, B, C)$ and (A, B, C) are equal and so, the characteristic subspaces $\hat{\mathcal{S}}(\lambda)$ of the triple $(\hat{A}, \hat{B}, \hat{C})$ are given by the dual counterpart of Lemma 2.7:

$$\hat{S}(\lambda) = \mathcal{S}(\lambda) \cap \mathcal{S}. \tag{13}$$

An admissible vector will be searched in the subspace $\hat{S}(\lambda)$.

LEMMA 3.4. *There exists a finite set* $\Phi \subset \mathbb{R}$ *with* $\operatorname{Card} \Phi \leqq n-p$ *such that for all* $\lambda \in \mathbb{C}\backslash\mathbb{R}$, *for all* $s \in \hat{\mathcal{S}}(\lambda)$, *we have*

$$\textit{for all } \beta \in \mathbb{C}\backslash\Phi, \quad \langle s, \bar{s}\rangle \cap \mathcal{T}(\beta) = 0,$$

$$\textit{for all } \beta \in \Phi, \quad \dim \langle s, \bar{s}\rangle \cap \mathcal{T}(\beta) = 1.$$

Proof. From Theorem 2.6, as $\mathcal{S} \cap \operatorname{Ker} C = 0$ and the fact that the elements of Λ'_F are distinct, we have

$$\text{for all } \beta \in \mathbb{C}\backslash\Lambda'_F, \quad \langle s, \bar{s}\rangle \cap \mathcal{T}(\beta) = 0, \tag{14}$$

$$\text{for all } \beta \in \Lambda'_F, \quad \dim \langle s, \bar{s}\rangle \cap \mathcal{T}(\beta) \leqq 1. \tag{15}$$

As $\hat{\mathcal{S}}(\lambda) \subset \mathcal{S}$ and $\mathcal{S} = \bar{\mathcal{S}}$, the lemma is proved (necessarily $\Phi \subset \Lambda'_F$). □

The proof of the following technical lemma is straightforward by using the Brunovski canonical form.

LEMMA 3.5. (i) *Assume that* $n' \geqq 3$, $m' \geqq 2$. Let $\theta \in \mathbb{C}\backslash\mathbb{R}$ *and* s_1, s_2 *be two independent vectors in* $\hat{S}(\theta)$. *Then, for almost any vector* $s_3 \in \hat{\mathcal{S}}(\theta)$, *we have*

$$\langle s_1, \bar{s}_1\rangle \cap \langle s_2, \bar{s}_2\rangle \neq \langle s_1, \bar{s}_1\rangle \cap \langle s_3, \bar{s}_3\rangle.$$

(ii) *Assume that* $n' \geqq 4$, $m'=1$. *Let* $\theta_1 \in \mathbb{C}\backslash\mathbb{R}$. *Then, for all* $\theta_2 \in \mathbb{C}\backslash(\mathbb{R} \cup \{\theta_1, \bar{\theta}_1\})$, *for all* $s_1 \in \hat{\mathcal{S}}(\theta_1)$, *for all* $s_2 \in \hat{\mathcal{S}}(\theta_2) (s_1 \neq 0, s_2 \neq 0)$, *we have*

$$\langle s_1, \bar{s}_1\rangle \cap \langle s_2, \bar{s}_2\rangle = 0.$$

(iii) *Assume that* $n'=3$, $m'=1$. *Let* $\theta_1, \theta_2 \in \mathbb{C}\backslash\mathbb{R}$ *such that* $\theta_2 \notin \{\theta_1, \bar{\theta}_1\}$. *Then, for any* $\theta_3 \in \mathbb{C}\backslash\mathbb{R}$ *such that* $\theta_3 \notin \{\theta_1, \bar{\theta}_1, \theta_2, \bar{\theta}_2\}$, *for any* $s_i \in \hat{\mathcal{S}}(\theta_i)$ $(i=1,2,3)$ *such that* $s_i \neq 0$, *we have*

$$\langle s_1, \bar{s}_1\rangle \cap \langle s_2, \bar{s}_2\rangle \neq \langle s_1, \bar{s}_1\rangle \cap \langle s_3, \bar{s}_3\rangle.$$

The following lemma gives more information about the number of pathological λ for a given β in the set Φ of Lemma 3.4.

LEMMA 3.6. *For any $\beta \in \Phi$, there exists a finite subset $\Xi(\beta)$ of $\mathbb{C}\backslash\mathbb{R}$ satisfying*

- *If $n' \geqq 4$,* Card $\Xi(\beta) \leqq 2$;
- *If $n' = 3$,* Card $\Xi(\beta) \leqq 4$;
- *For any $\lambda \notin \Xi(\beta)$, for almost any $s \in \hat{\mathscr{S}}(\lambda)$, $\langle s, \bar{s}\rangle \cap \mathscr{T}(\beta) = 0$.*

Proof. Let $\beta \in \Phi$. Three cases are to be considered.

(a) $n' \geqq 3$; $m' \geqq 2$. Let $\theta \in \mathbb{C}\backslash\mathbb{R}$; assume that there exist two independent vectors $s_1, s_2 \in \hat{\mathscr{S}}(\theta)$ such that $\langle s_1, \bar{s}_1\rangle \cap \mathscr{T}(\beta) \neq 0$ and $\langle s_2, \bar{s}_2\rangle \cap \mathscr{T}(\beta) \neq 0$. Then, from Lemma 3.4, $\dim \langle s_1, \bar{s}_1\rangle \cap \mathscr{T}(\beta) = 1$ and $\dim \langle s_2, \bar{s}_2\rangle \cap \mathscr{T}(\beta) = 1$. From (15), this means that $\langle s_1, \bar{s}_1\rangle \cap \mathscr{T}(\beta) = \langle s_2, \bar{s}_2\rangle \cap \mathscr{T}(\beta)$, which is equivalent to $\langle s_1, \bar{s}_1\rangle \cap \langle s_2, \bar{s}_2\rangle \cap \mathscr{T}(\beta) \neq 0$. So, from Lemma 3.5(i), for almost all vectors $s_3 \in \hat{\mathscr{S}}(\theta)$, we have $\langle s_3, \bar{s}_3\rangle \cap \mathscr{T}(\beta) = 0$ (otherwise we would have $\langle s_1, \bar{s}_1\rangle \cap \mathscr{T}(\beta) \neq \langle s_3, \bar{s}_3\rangle \cap \mathscr{T}(\beta)$, that is $\dim \mathscr{S} \cap \mathscr{T}(\beta) = 2$). So, in this case, Card $\Xi(\beta) = 0$.

(b) $n' \geqq 4$; $m' = 1$. In this case, the above argument used with Lemma 3.5(ii) leads to Card $\Xi(\beta) \leqq 2$.

(c) $n' = 3$; $m' = 1$. In this case, from Lemma 3.5(iii), Card $\Xi(\beta) \leqq 4$. □

The previous technical results can be summarized in the following lemma.

LEMMA 3.7. *There exists a finite subset $\Xi \in \mathbb{C}\backslash\mathbb{R}$ such that*

- *If $n' = p \geqq 4$,* Card $\Xi \leqq 2(n-p)$;
- *If $p = 3$,* Card $\Xi \leqq 4(n-p)$;
- *For any $\lambda \in \mathbb{C}\backslash\mathbb{R} \cup \Xi$, there exists $s \in \mathscr{S}(\lambda)$ such that for any $\beta \in \mathbb{C}$, $\langle s, \bar{s}\rangle \cap \mathscr{T}(\beta) = 0$.*

Proof. From Lemmas 3.4 and 3.6, we can take $\Xi = \bigcup_{\beta \in \Phi} \Xi(\beta)$ and choose s generically in $\hat{\mathscr{S}}(\lambda)$ for any $\lambda \notin \Xi$. □

Step 2. Genericity of admissible vectors. We will show now that if there exists one admissible vector in $\mathscr{S}(\lambda)$, almost every vector in $\mathscr{S}(\lambda)$ will be admissible as well. The following supporting lemma will be used.

LEMMA 3.8. *Let $P_1(x, \beta), \cdots, P_n(x, \beta)$ be polynomials in the complex variables $x \in \mathbb{C}^m$ and $\beta \in \mathbb{C}$ such that there exists x_0 satisfying*

$$\textit{for any } \beta \in \mathbb{C}, \textit{ there exists } i \in [1, n] \textit{ such that } P_i(x_0, \beta) \neq 0.$$

Then, the algebraic set

$$\mathscr{V} = \{x \textit{ s.t. there exists } \beta \textit{ s.t. } P_1(x, \beta) = 0, \cdots, P_n(x, \beta) = 0\}$$

is thin in $\mathbb{C}^m$.

Proof. If at least one of the considered polynomials is constant with respect to β but is not identically equal to zero, it is clear that $\mathscr{V}$ is thin. Otherwise, it is clear that there exists at least two polynomials that depend on β. Assume that P_1 is such a polynomial. So its leading coefficient is not identically equal to zero ($Ldc_\beta P_1(x) \neq 0$). It is straightforward to construct another polynomial $P_1'(x, \beta)$ as a linear combination of $P_2, \cdots, P_n$ such that $P_1(x_0, \beta)$ and $P_1'(x_0, \beta)$ have no common root and with $Ldc_\beta P_1'(x) \neq 0$. By a continuity argument, the property relative to the common roots is true in a neighbourhood of x_0. As the leading coefficients are not identically equal to zero, there exists x_1 arbitrarily close to x_0 satisfying:

- $P_1(x_1, \beta)$ and $P_1'(x_1, \beta)$ have no common root,
- $Ldc_\beta(P_1(x_1)) \neq 0$,

- $Ldc_\beta(P_1'(x_1)) \neq 0$,

i.e., $Res_\beta(P_1, P_1')(x_1) \neq 0$. So, $Res_\beta(P_1, P_1')(x) \neq 0$. Therefore, as

$$\mathcal{V} \subset \{x \text{ s.t. } Res_\beta(P_1, P_1')(x) = 0\}, \ \mathcal{V} \text{ is thin.} \qquad \square$$

The fact that admissibility is a generic condition can now be proved.

LEMMA 3.9. *Let $\lambda \in \mathbb{C}\backslash(\mathbb{R} \cap \Xi)$. Then, for almost every $s \in \mathcal{S}(\lambda)$, we have*

$$\textit{for all } \beta \in \mathbb{C}, \quad \langle s, \bar{s}\rangle \cap \mathcal{T}(\beta) = 0.$$

Proof. Let $\mathcal{T}(\beta)$ be the rational matrix such that $\mathcal{T}(\beta) = \operatorname{Im} T(\beta)$. Let λ be a given complex number in $\mathbb{C}\backslash(\mathbb{R} \cap \Xi)$ and S be a matrix such that $\mathcal{S}(\lambda) = \operatorname{Im} S$. We have to prove that the set

$$\mathcal{V}(\lambda) = \{x \in \mathbb{C}^m \text{ such that there exists } \beta \in \mathbb{C} \text{ s.t. } \langle Sx, \overline{Sx}\rangle \cap \mathcal{T}(\beta) \neq 0\}$$

is thin in $\mathbb{C}^m$. We have

$$x \in \mathcal{V}(\lambda) \text{ iff there exists } \beta \in \mathbb{C} \text{ s.t. rank}\,[Sx\,\overline{Sx}\;\, T(\beta)] \leqq n - p + 1.$$

Let $(P_i)_{i \in I}$ be the minors of order $n-p+2$ of the matrix $[Sx\,\overline{Sx}\;\, T(\beta)]$. Clearly, the P_i's are polynomials in x, $\bar{x}$, β and

$$x \in \mathcal{V}(\lambda) \text{ iff there exists } \beta \in \mathbb{C} \text{ s.t. } P_i(x, \bar{x}, \beta) = 0.$$

From Lemma 3.7, there exists at least one admissible vector, i.e.,

$$\text{there exists } x_0 \in \mathbb{C}^m \text{ such that for all } \beta \in \mathbb{C}, \text{ there exists } i \in I \text{ s.t. } P_i(x_0, \bar{x}_0, \beta) \neq 0.$$

Lemma 3.8 concludes the proof. □

Step 3. *End of the proof of Proposition* 3.3. Consider the restricted system constructed in Step 1. By construction, for any $s \in \hat{\mathcal{S}}(\lambda) \cap \mathcal{S}$, we have $\langle s, \bar{s}\rangle \cap \operatorname{Ker} C = 0$. So,

$$\dim(\operatorname{Ker} C + \mathcal{S}(\lambda)) \geqq \dim \operatorname{Ker} C + 1,$$

$$\dim(\operatorname{Ker} C + \mathcal{S}(\lambda) + \overline{\mathcal{S}(\lambda)}) \geqq \dim \operatorname{Ker} C + 2.$$

From the Kimura lemma (see Lemma 1 of [7]), for almost any vector $s \in \mathcal{S}(\lambda)$, we have that $\langle s, \bar{s}\rangle \cap \operatorname{Ker} C = 0$. Furthermore, using the Brunovski canonical form, it is clear that almost any vector $s \in \mathcal{S}(\lambda)$ satisfies the properties (11) and (12). So, the properties (8) to (12) are independently satisfied by almost any vector $s \in \mathcal{S}(\lambda)$. As a finite union of thin subsets is thin, almost every vector of $\mathcal{S}(\lambda)$ satisfies all of them. □

3.3. Recursive construction of the lattice. To construct $\mathcal{S}_{i+1}$ from $\mathcal{S}_i$, the proposed approach consists in applying the results of the previous section to the subsystem induced modulo $\mathcal{S}_i$. The genericity of the properties obtained relative to this subsystem will be recovered for the subsystem induced modulo $\mathcal{S}_{i+1}$ and so on.

PROPOSITION 3.10. *Let (A, B, C) be a complete triple satisfying $m + p > n$. Consider an (A, B)-invariant subspace $\mathcal{S}_i$ with the following properties.*

(i) *There exists a self-conjugate set of distinct complex numbers $\{\lambda_1, \cdots, \lambda_i\}$ such that*

$$\mathcal{S}_i = \langle s_1, \cdots, s_i\rangle \textit{ with } s_i \in \mathcal{S}(\lambda_i) \textit{ and } s_k = \bar{s}_j \textit{ if } \lambda_k = \bar{\lambda}_j;$$

(ii) $\mathcal{S}_i \cap \operatorname{Ker} C = 0$;

(iii) $\mathcal{S}_i \cap \mathcal{T}(\beta) = 0$ *for any* $\beta \in \mathbb{C}$;

(iv) $\mathcal{S}_i \cap Im\, B = 0$ *if* $i \leqq n - m$.

Let λ_i be a real number (respectively, a complex nonreal number) and assume that $i \leqq p - 1$ (respectively, $i < p - 1$). Then, there exists a finite set $\Xi_{i+1} \subset \mathbb{C}$ (cf. Propositions 3.2 *and*

3.3 *for a majoration of* Card Ξ_{i+1}) *such that if* $\lambda_{i+1} \notin \{\lambda_1, \cdots, \lambda_i\} \cup \Xi_{i+1}$, *then, for almost any* $s_{i+1} \in \mathcal{S}(\lambda_{i+1})$, *the subspace* $\mathcal{S}_{i+1} = \langle s_1, \cdots, s_i, s_{i+1}\rangle$ (*respectively,* $\mathcal{S}_{i+2} = \langle s_1, \cdots, s_i, s_{i+1}, \bar{s}_{i+1}\rangle$) *satisfies the same properties* (i)–(iv) *than* $\mathcal{S}_i$. *Furthermore,* $\Xi_i \subset \Xi_{i+1}$.

Proof. This result will be proved only in the case where λ_{i+1} is real, the demonstration in the complex case being identical. Let $\mathcal{S}_i$ be a subspace satisfying the properties (i)–(iv). As $\mathcal{S}_i \cap \operatorname{Ker} C = 0$, there exists an output feedback gain K_i such that $\mathcal{S}_i$ is $(A + BK_iC)$-invariant. Let us consider the system $(\tilde{A}, \tilde{B}, \tilde{C})$ induced by $(A + BK_iC, B, C)$ modulo $\mathcal{S}_i$. Let $\tilde{\mathcal{S}}(\lambda)$ and $\tilde{\mathcal{T}}(\beta)$ denote the characteristic subspaces of the induced system. It is well known that the pair $(\tilde{A}, \tilde{B})$ is controllable (cf. [17]). As $\mathcal{S}_i \cap \operatorname{Ker} C = 0$ and $\mathcal{S}_i \cap \mathcal{T}(\beta) = 0$ for any $\beta \in \mathbb{C}$, the pair $(\tilde{A}, \tilde{C})$ is observable (cf. Comment 1). Furthermore, if $\tilde{n} \triangleq \dim \mathcal{X}/\mathcal{S}_i$, $\tilde{p} \triangleq \dim \operatorname{Im} \tilde{C}$, $\tilde{m} \triangleq \dim \operatorname{Im} \tilde{B}$, we have

$$\begin{aligned}
&\tilde{n} = n - i,\\
&\tilde{p} = p - i \quad (\text{as } \mathcal{S}_i \cap \operatorname{Ker} C = 0),\\
&\tilde{m} = m \quad \text{for } i \leqq n - m \quad (\text{as then } \mathcal{S}_i \cap \operatorname{Im} B = 0),\\
&\quad = n - i \quad \text{for } i > n - m.
\end{aligned} \tag{16}$$

So, in any case, the induced system satisfies the condition $\tilde{n} < \tilde{m} + \tilde{p}$ and Proposition 3.2 may be applied to the induced system. Thus, there exists no more than one real number $\tilde{\lambda}$ such that for any $\lambda_{i+1} \in \mathbb{R}\backslash\{\tilde{\lambda}\}$, for almost any vector $\tilde{s}_{i+1} \in \tilde{\mathcal{S}}(\lambda_{i+1})$, we have

$$\begin{aligned}
&\langle \tilde{s}_{i+1}\rangle \cap \operatorname{Ker} \tilde{C} = 0,\\
&\langle \tilde{s}_{i+1}\rangle \cap \tilde{\mathcal{T}}(\beta) = 0 \quad \text{for all } \beta \in \mathbb{C},\\
&\langle \tilde{s}_{i+1}\rangle \cap \operatorname{Im} \tilde{B} = 0 \quad \text{if } i < n - m.
\end{aligned} \tag{17}$$

In the sequel, we will assume that $\lambda_{i+1} \notin \{\lambda_1, \cdots, \lambda_i\} \cup \{\tilde{\lambda}\}$. The two following lemmas will achieve the proof of the proposition.

LEMMA 3.11. *For almost any vector* $s_{i+1} \in \mathcal{S}(\lambda_{i+1})$, *the subspace* $\mathcal{S}_{i+1} = \mathcal{S}_i + \langle s_{i+1}\rangle$ *satisfies the conditions* (i)–(iv) *of Proposition* 3.10.

Proof. From Lemma 2.7, we have

$$\pi(\operatorname{Im} B) = \operatorname{Im} \tilde{B}, \tag{18}$$

$$\pi(\operatorname{Ker} C) = \operatorname{Ker} \tilde{C}, \tag{19}$$

$$\pi(\mathcal{S}(\lambda)) = \tilde{\mathcal{S}}(\lambda) \quad \text{for all } \lambda \notin \{\lambda_1, \cdots, \lambda_i\}, \tag{20}$$

$$\pi(\mathcal{T}(\beta)) = \tilde{\mathcal{T}}(\beta) \quad \text{for all } \beta \in \mathbb{C}. \tag{21}$$

Now, consider a vector $\tilde{s}_{i+1} \in \tilde{\mathcal{S}}(\lambda_{i+1})$ satisfying (17). From (20), we can associate to $\tilde{s}_{i+1}$ a vector $s_{i+1} \in \mathcal{S}(\lambda_{i+1})$ such that $\pi(\mathcal{S}_{i+1}) = \langle \tilde{s}_{i+1}\rangle$. Using the fact that if $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ *are three subspaces in a space* $\mathcal{X}$ *such that* $\mathcal{B} \subset \mathcal{A}$, $\mathcal{B} \cap \mathcal{C} = 0$ *and* $\pi(\mathcal{A}) \cap \pi(\mathcal{C}) = 0$, *where* π *is the canonical surjection from* $\mathcal{X}$ *onto* $\mathcal{X}/\mathcal{B}$, *then* $\mathcal{A} \cap \mathcal{C} = 0$. Equations (18)–(21), $\mathcal{S}_{i+1}$ satisfy properties (i)–(iv) of the proposition. Furthermore, the genericity of $\tilde{s}_{i+1}$ in $\tilde{\mathcal{S}}(\lambda_{i+1})$ corresponds obviously to the genericity of s_{i+1} in $\mathcal{S}(\lambda_{i+1})$. □

LEMMA 3.12. *Let* Ξ_k *be the set of the eigenvalues which cannot be assigned at the kth step of the algorithm. Then* $\Xi_i \subset \Xi_{i+1}$ *for* $i = 1, \cdots, p - 2$.

Proof. Assume that λ is a "pathological" eigenvalue at the first step of the algorithm. This means that for any $s \in \mathcal{S}(\lambda)$, there exists $\beta \in \mathbb{C}$ such that $\langle s, \bar{s}\rangle \cap \mathcal{T}(\beta) \neq 0$. Let π be the projection from $\mathcal{X}$ onto $\mathcal{X}/\mathcal{S}_1$ and let $\tilde{s} \triangleq \pi(s)$. In view of (21), we have

$$\pi(\langle s, \bar{s}\rangle \cap \mathcal{T}(\beta)) \subset \langle \tilde{s}, \bar{\tilde{s}}\rangle \cap \tilde{\mathcal{T}}(\beta).$$

As $\mathcal{S}_1 \cap \mathcal{T}(\beta) = 0$ by construction, we have

$$\langle s, \bar{s}\rangle \cap \mathcal{T}(\beta) \neq 0 \Rightarrow \langle \tilde{s}, \bar{\tilde{s}}\rangle \cap \tilde{\mathcal{T}}(\beta) \neq 0.$$

So, if λ is pathological at the first step, it is still pathological at the second step. Therefore it remains pathological all along the algorithm. □

Note that the subspaces $\mathcal{S}_i$, $i = 1, \cdots, p-1$ so constructed from a generic choice of vectors $s_i \in \mathcal{S}(\lambda_i)$ satisfy the lattice given in Fig. 3.1:

1. $\mathcal{S}_i$ is (A, B)-invariant (from (i));
2. $\mathcal{S}_i$ is (C, A)-invariant (from (ii));
3. $\mathcal{N}^*_{\mathcal{S}_i} = \mathcal{S}_i$ (from (iii) and Comment 1);
4. $\mathcal{R}^*_{\mathcal{S}_i} = 0$ if $i \leqq n - m$ (from (iv));
5. By construction, the eigenvalues are assigned as desired.

3.4. Final step of the construction of the lattice. Now, the subspace $\mathcal{S}_p$ of the lattice must be constructed. In the following proposition are pointed out the degrees of freedom appearing in this construction.

PROPOSITION 3.13. *Consider a subspace* $\mathcal{S}_{p-1} = \langle s_1, \cdots, s_{p-1}\rangle$ *satisfying conditions* (i)–(iv) *of Proposition* 3.10. *Let* λ_p *be a real number and let* $\Lambda_2 = \{\lambda_{p+1}, \cdots, \lambda_n\}$ *be a self-conjugate set of complex numbers, so that* $\Lambda = \{\lambda_1, \cdots, \lambda_n\}$ *is admissible. Then the subspace*

$$\mathcal{S}'(\lambda_p) = \mathcal{S}(\lambda_p) \cap (\mathcal{T}(\lambda_{p+1}) \oplus \mathcal{S}_{p-1}) \cap \cdots \cap (\mathcal{T}(\lambda_n) \oplus \mathcal{S}_{p-1})$$

satisfies

1. $\dim \mathcal{S}'(\lambda_p) = m + p - n > 0$;
2. *For almost every vector* $s_p \in \mathcal{S}'(\lambda_p)$, *the subspace* $\mathcal{S}_p = \langle s_1, \cdots, s_p\rangle$ *satisfies* $\mathcal{S}_p \cap \operatorname{Ker} C = 0$ *and the spectrum induced modulo* $\mathcal{S}_p$ *is* Λ_2.

Proof. Consider the system induced modulo $\mathcal{S}_{p-1}$ (the related canonical surjection is denoted by π). For this system, we have (cf. (16)) $\tilde{n} = n - p + 1$, $\tilde{m} = n - p + 1$, $\tilde{p} = 1$. Thus, for all $\lambda \in \mathbb{C}$, $\tilde{\mathcal{S}}(\lambda) = \tilde{\mathcal{X}}$. Furthermore, as $\mathcal{S}_{p-1} \cap \operatorname{Ker} C = 0$ and $\lambda_p \notin \{\lambda_1, \cdots, \lambda_{p-1}\}$, the relations (18) to (21) are satisfied. So, as $\operatorname{Ker} \pi = \mathcal{S}_{p-1}$, we have

$$\pi(\mathcal{S}'(\lambda_p)) = \tilde{\mathcal{T}}(\lambda_{p+1}) \cap \cdots \cap \tilde{\mathcal{T}}(\lambda_n). \tag{22}$$

Considering the special form of the characteristic subspaces on the Brunovski canonical form, it is straightforward to see that $\dim \tilde{\mathcal{T}}(\lambda_{p+1}) \cap \cdots \cap \tilde{\mathcal{T}}(\lambda_n) = 1$ and $\tilde{\mathcal{T}}(\lambda_{p+1}) \cap \cdots \cap \tilde{\mathcal{T}}(\lambda_n) \cap \operatorname{Ker} \tilde{C} = 0$. Therefore from (22),

$$\dim \pi(\mathcal{S}'(\lambda_p)) = 1 \tag{23}$$

and

$$\pi(\mathcal{S}'(\lambda_p) \cap (\operatorname{Ker} C \oplus \mathcal{S}_{p-1})) = 0. \tag{24}$$

Furthermore, as $\mathcal{S}(\lambda_p) \cap \mathcal{S}_{p-1} = \operatorname{Ker} \pi | \mathcal{S}(\lambda_p)$, we have

$$\dim \mathcal{S}(\lambda_p) \cap \mathcal{S}_{p-1} = \dim \mathcal{S}(\lambda_p) - \dim \operatorname{Im} \pi | \mathcal{S}(\lambda_p) \tag{25}$$

$$= \dim \mathcal{S}(\lambda_p) - \dim \tilde{\mathcal{X}} \tag{26}$$

so

$$\dim \mathcal{S}(\lambda_p) \cap \mathcal{S}_{p-1} = m + p - n - 1. \tag{27}$$

From (23) and the fact that $\mathcal{S}'(\lambda_p) \cap \mathcal{S}_{p-1} = \mathcal{S}(\lambda_p) \cap \mathcal{S}_{p-1}$, we have

$$\dim \mathcal{S}'(\lambda_p) = \dim \mathcal{S}'(\lambda_p) \cap \mathcal{S}_{p-1} + \dim \pi(\mathcal{S}'(\lambda_p)) = m + p - n.$$

Hence (i) holds. Furthermore, (24) leads to

$$\mathcal{S}'(\lambda_p)\cap(\operatorname{Ker} C\oplus\mathcal{S}_{p-1})\subset\mathcal{S}_{p-1}.$$

Therefore, $\mathcal{S}'(\lambda_p)\cap(\operatorname{Ker} C\oplus\mathcal{S}_{p-1})$ is contained in $\mathcal{S}(\lambda_p)\cap\mathcal{S}_{p-1}$. As this subspace is a hyperplane of $\mathcal{S}'(\lambda_p)$ (from (27)), we deduce that for almost any vector $s_p\in\mathcal{S}'(\lambda_p)$, the subspace $\mathcal{S}_p=\langle s_1,\cdots,s_p\rangle$ satisfies $\mathcal{S}_p\cap\operatorname{Ker} C=0$. Moreover, by construction, $\mathcal{S}_p\cap\mathcal{T}(\lambda_i)\neq 0$ for $i=p+1,\cdots,n$. From Theorem 2.6, this means that the spectrum induced modulo $\mathcal{S}_p$ is $\{\lambda_{p+1},\cdots,\lambda_n\}$.

The construction of the subspace $\mathcal{S}_p=\langle s_1,\cdots,s_p\rangle$ from a generic choice of vectors $s_i\in\mathcal{S}(\lambda_i)$, $i=1,\cdots,p-1$ and $s_p\in\mathcal{S}'(\lambda_p)$ completes that of the lattice given in Fig. 3.1. So, Theorem 3.1 is demonstrated. □

3.5. Illustrative example. Let us consider the following numerical example:

$$A=\begin{bmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\\0&0&0&0\end{bmatrix},\qquad B=\begin{bmatrix}2&1\\5&3\\6&4\\2&2\end{bmatrix},$$

$$C=\begin{bmatrix}1&0&0&0\\0&1&0&0\\0&0&1&0\end{bmatrix}$$

(i) *The generic case.* Assume that $\{\lambda_1=-2+2i,\ \lambda_2=-2-2i,\ \lambda_3=-2,\ \lambda_4=-1\}$. We have to choose a vector in $\mathcal{S}(-2+2i)$. It is straightforward to check that

$$\mathcal{S}(-2+2i)=\begin{bmatrix}5&0\\0&1\\-12+4i&2\\-28-4i&2\end{bmatrix}.$$

Assume that $s_1=[5+5i,10,4-8i,-4-32i]^T$. The vector s_2 is fixed as being equal to $\bar{s}_1$. It remains to find s_3 in the subspace $\mathcal{S}'(-2)=\mathcal{S}(-2)\cap(\mathcal{T}(-1)+\langle s_1,s_2\rangle)$. As $\mathcal{T}(-1)=\operatorname{Im}[0,0,1,1]^T$, we obtain $s_3=[-1,-0.5,-1,3]^T$. Now the input directions w_i are given by solving $(A-\lambda_i I)s_i=-Bw_i$ for $i=1,3$. The following numerical values are obtained:

$$w_1=\begin{bmatrix}-66-28i\\102+56i\end{bmatrix},\qquad w_3=\begin{bmatrix}5.5\\-8.5\end{bmatrix}.$$

Finally, the feedback gain is given by $K=[w_1,\bar{w}_1,w_3](C[s_1,\bar{s}_1,s_3])^{-1}$:

$$K=\begin{bmatrix}-4&-5&1\\7.2&7.6&-2.5\end{bmatrix}.$$

We are now going to illustrate the pathologies that may occur (but which are avoided by choosing generic eigenvalues and eigenvectors).

(ii) *First class of pathologies.* Assume that the same eigenvalues are to be assigned. If the vector s_1 is selected as being equal to $[5,0,-12+4i,-28-4i]^T$, then as $\mathcal{T}(\beta)=\operatorname{Im}[0,0,1,-\beta]^T$, there exists obviously a value of β for which $\langle s_1,\bar{s}_1\rangle\cap\mathcal{T}(\beta)\neq 0$ (this value is $\beta=1$). It is the pathology corresponding to equation (10). Now if s_1 is selected as being equal to $[5,-2,-12,-28-8i]^T$, clearly $\langle s_1,\bar{s}_1\rangle\cap\operatorname{Ker} C\neq 0$ (see (9)). Note that it is clear that almost any other choice of s_1 does not induce such pathologies provided that the vector of the first and the second entries of s_1 is not colinear to a real vector.

(iii) *Second class of pathologies.* This numerical example has been especially worked out in order to exhibit another kind of pathology which does not appear generally. If $\lambda_1 = -1+i$, we have

$$\mathcal{S}(-1+i)=\begin{bmatrix} 0 & 0 \\ 1-i & 1 \\ 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

Clearly, for all choice of s_1 in this subspace, it is clear that there exists a value of β denoted $\beta(s_1)$ such that $\langle s_1, \overline{s_1}\rangle \cap \mathcal{T}(\beta(s_1)) \neq 0$. The set of such complex numbers was denoted Ξ. Its cardinal was only estimated from above in Proposition 3.3, this numerical example shows that it is not empty for some system. (Note that similar pathologies have already been pointed out [8], [10] but only for real numbers.)

4. A necessary and sufficient condition for exact pole assignment by output feedback. In the case where $m+p>n$, Kimura [7] and Davison and Wang [2] have shown that it was possible by using output feedback to assign the poles of a complete triple arbitrarily close to any prescribed set of eigenvalues. The algorithm developed in § 3 also deals with this kind of approximative assignment. Nevertheless, our results are much more precise since we have shown that only a *finite* number of eigenvalues cannot be assigned (and that it is possible to assign up to $p-1$ eigenvectors arbitrarily closed to prescribed vectors in (A, B)-characteristic subspaces). The preciseness of Propositions 3.2 and 3.3 can be used, furthermore, to improve these results by deriving a shortened proof of the necessary and sufficient condition for *exact* pole assignment given in Magni [9], [10]. At each step of the algorithm used in the previous section, the eigenstructure assignment was relative to *induced* subsystems (see Fig. 3.1). Here, as we are not concerned any more with eigenvector assignment, we have, as an additional freedom, the possibility of considering at each step, either *induced* or *restricted* subsystems. It is this additional freedom which permits us to cope with exact assignability. Let us recall the main result of [9] or [10].

THEOREM 4.1. *Let (A, B, C) be a complete triple satisfying $m+p>n$. A self-conjugate set Λ of n distinct eigenvalues being given, here exists an output feedback gain K such that $\sigma(A+BKC)=\Lambda$ if and only if the three following conditions do not hold simultaneously*:

1. $n=m+p-1$, *p and m are even*;
2. *Λ contains exactly one real eigenvalue denoted by $\tilde{\lambda}$*;
3. $\mathcal{C}_0=\mathcal{S}(\tilde{\lambda})$ *or* $\mathcal{B}_0=\mathcal{T}(\tilde{\lambda})$.

Note that the real number denoted $\tilde{\lambda}$ in Theorem 4.1 is the one defined by Lemma 2.3 and that both conditions $\mathcal{C}_0=\mathcal{S}(\tilde{\lambda})$ and $\mathcal{B}_0=\mathcal{T}(\tilde{\lambda})$ are equivalent (see [8]). The proof of the necessity of this theorem can be found in [8]. The sufficiency was proved in two steps. The first step [3], [6] is relative to the case where $\Lambda \cap \mathbb{R} = \emptyset$. The second step deals with the general case [9], [10]. The preciseness of Proposition 3.3 allows us to derive shortened proof of the first step. Let us state the result which corresponds to this first step.

PROPOSITION 4.2. *We have a complete triple (A, B, C) such that $m+p>n$ with n even. For all self-conjugate set Λ of n distinct complex nonreal numbers, there exists a gain output feedback K such that $\sigma(A+BKC)=\Lambda$.*

This result will be proved by downwards recursion. The following lemma concerns the final step of the recursion.

LEMMA 4.3. *If $m+p>n$ with $m \leqq 3$ and $p \leqq 3$, Proposition 4.2 is true.*

Proof. If $m=n$ or $p=n$, we can conclude by using state feedback or output injection. So, we have only to consider the following cases (n being even):

1. $m=3$, $p=3$, $n=4$;
2. $m=2$, $p=3$, $n=4$ with controllability indices $\{2,2\}$;
3. $m=2$, $p=3$, $n=4$ with controllability indices $\{3,1\}$.

First case. The sets of controllability and observability indices are both $\{2,1,1\}$. Considering the Brunovski canonical form for $\mathcal{S}(\lambda)$, we have (see § 2)

$$\dim \mathcal{C}_0=2,\quad \dim \mathcal{S}(\lambda)=3,\quad \dim \mathcal{S}(\lambda)+\mathcal{S}(\bar{\lambda})=4.$$

So, from the Kimura lemma [7, Lemma 1], almost all vector $s\in\mathcal{S}(\lambda)$ will satisfy

$$\langle s,\bar{s}\rangle\cap\mathcal{C}_0=0, \tag{28}$$

hence from (1)

$$\langle s,\bar{s}\rangle\cap\mathcal{T}(\beta)=0\quad\text{for all }\beta\in\bar{\mathbb{C}}.$$

Furthermore, from the fact that $\mathcal{S}(\lambda)\neq \operatorname{Im} B$, it is clear that almost all vector $s\in\mathcal{S}(\lambda)$ satisfies

$$\dim(\langle s,\bar{s}\rangle+\operatorname{Im} B)=4. \tag{29}$$

So, from Comment 1, the system $(A(\operatorname{mod}\langle s,\bar{s}\rangle), B(\operatorname{mod}\langle s,\bar{s}\rangle), C(\operatorname{mod}\langle s,\bar{s}\rangle))$ is controllable and observable. As it has two states, one output, and two inputs, it is pole assignable by output injection. □

Second case. The observability indices being $\{2,1,1\}$, we have

$$\dim \mathcal{C}_0=2.$$

If $\mathcal{S}(\lambda)\subset\mathcal{C}_0$, we have $\mathcal{S}(\bar{\lambda})\subset\mathcal{C}_0$, hence $\mathcal{S}(\lambda)+\mathcal{S}(\bar{\lambda})\subset\mathcal{C}_0$, which is nonsense since $\dim\mathcal{S}(\lambda)+\mathcal{S}(\bar{\lambda})=4$ (obvious considering the Brunovsky canonical form). So, $\dim\mathcal{C}_0+\mathcal{S}(\lambda)\geqq 3$ and $\dim\mathcal{C}_0+\mathcal{S}(\lambda)+\mathcal{S}(\bar{\lambda})=4$. From the Kimura lemma [7, Lemma 1], it follows that almost any $s\in\mathcal{S}(\lambda)$ satisfies (28). Furthermore, it is clear that $\dim\operatorname{Im} B+\mathcal{S}(\lambda)=3$ and $\dim\operatorname{Im} B+\mathcal{S}(\lambda)+\mathcal{S}(\bar{\lambda})=4$, therefore, as above, almost any $s\in\mathcal{S}(\lambda)$ satisfies (29). Considering the subsystem induces modulo $\langle s,\bar{s}\rangle$, we can conclude as above. □

Third case. To conclude as in the above case, it suffices to prove that

$$\dim(\mathcal{C}_0+\mathcal{S}(\lambda)+\mathcal{S}(\bar{\lambda}))=4.$$

If this condition is not satisfied, we have

$$\mathcal{C}_0\subset\mathcal{S}(\lambda)+\mathcal{S}(\bar{\lambda})$$

because $\dim(\mathcal{S}(\lambda)+\mathcal{S}(\bar{\lambda}))=3$ (still considering the Brunovski canonical form). Let λ' be another eigenvalue of Λ, so, $\Lambda=\{\lambda,\bar{\lambda},\lambda',\bar{\lambda}'\}$. Now, we will show that we cannot have simultaneously

$$\mathcal{C}_0\subset\mathcal{S}(\lambda)+\mathcal{S}(\bar{\lambda})\quad\text{and}\quad\mathcal{C}_0\subset\mathcal{S}(\lambda')+\mathcal{S}(\bar{\lambda}'),$$

which concludes the proof since either λ or λ' can be assigned. Now, assume that the above conditions are satisfied. Then

$$\mathcal{C}_0=(\mathcal{S}(\lambda)+\mathcal{S}(\bar{\lambda}))\cap(\mathcal{S}(\lambda')+\mathcal{S}(\bar{\lambda}')). \tag{30}$$

This relation will be written on the controllability canonical basis. In this basis, A, B, $\mathcal{S}(\lambda)$ take the form

$$(31)\qquad A=\begin{bmatrix}0&1&0&0\\0&0&1&0\\ *&*&*&*\\ *&*&*&*\end{bmatrix},\quad B=\begin{bmatrix}0&0\\0&0\\1&0\\0&1\end{bmatrix},\quad \mathcal{S}(\lambda)=\begin{bmatrix}1&0\\ \lambda&0\\ \lambda^2&0\\0&1\end{bmatrix}.$$

Let us define $x_1x_2x_3x_4$ as follows:

$$\operatorname{Ker} C \triangleq \operatorname{Im}[x_1, x_2, x_3, x_4]^T.$$

From (30), the vector $[0, 0, 0, 1]^T$ lies in $\mathscr{C}_0$. As $\mathscr{C}_0=\operatorname{Ker} C + A \operatorname{Ker} C$ and as $\dim \mathscr{C}_0=2$, we must have

$$\operatorname{rank}\begin{bmatrix}x_1&x_2&0\\x_2&x_3&0\\x_3&*&0\\x_4&*&1\end{bmatrix}=2.$$

If $x_1=0$, then necessarily $x_2=0$, hence $x_3=0$, so $\mathscr{C}_0=\operatorname{Im} B$ which, in view of (30) and (31) is a nonsense. If $x_1\neq 0$, take $x_1=1$ and $x_2=\varepsilon$. Then, $x_3=\varepsilon^2$ and $[1, \varepsilon, \varepsilon^2, 0]^T\in(\mathcal{S}(\lambda)+\mathcal{S}(\bar\lambda))\cap(\mathcal{S}(\lambda')+\mathcal{S}(\bar\lambda'))$. Hence, $\varepsilon=\lambda$ or $\bar\lambda$ and $\varepsilon=\lambda'$ or $\bar\lambda'$ which contradicts the fact that the elements of Λ are distinct. □

Proof of Proposition 4.2. The proof is performed by downwards recursion until conditions of Lemma 4.3 are encountered. Three initial configurations are to be considered.

Case 1. $p\geqq m$ and $p\geqq 4$. In this case, since n is even, $n<m+p\Rightarrow n<2p\Rightarrow n\leqq 2p-2\Rightarrow 2(n-p)\leqq n-2$. From Proposition 3.3, the number Ξ of complex eigenvalues which cannot be assigned satisfies Card $\Xi\leqq 2(n-p)$. So, there exists a complex number in Λ which can be assigned. Let s, $\bar s$ be the assigned right eigenvectors. At the next step, the system $(A(\text{mod}\langle s,\bar s\rangle), B(\text{mod}\langle s,\bar s\rangle), C(\text{mod}\langle s,\bar s\rangle))$ will be considered. This system has $n-2$ states, $p-2$ outputs, $\min(n-2, m)$ inputs. If $n-2=\min(n-2, m)$, the algorithm stops by the use of an output injection. Otherwise, it reduces to one of the three cases considered here.

Case 2. $m\geqq p$ and $m\geqq 4$. This is the dual case of the previous one. A pair of left eigenvectors is assigned; the next step is relative to a restricted system.

Case 3. $m\leqq 3$ and $p\leqq 3$. In this case, Lemma 4.3 can be used to conclude. □

5. Conclusion. In this paper we have studied the genericity of eigenvalue/eigenvector assignment for control systems in which the number of outputs (p) plus the number of inputs (m) exceeds the number of states (n). Intermediate results occurring in this proof have permitted us to obtain a necessary and sufficient condition for the exact assignability of a set of distinct poles. Maximal controllability subspaces and minimal complementary observability subspaces have been characterized in terms of the dimension of their intersections with the characteristic subspaces. It is this characterization that is the key to all results we have obtained here. It is worth noting that the main difference between the work presented in this paper and the work that is to be done for dealing with more general systems $(m+p\leqq n)$ lies in the fact that here (except in the proof of Theorem 4.1 relative to a very special case which is detailed in [10]) it was possible to construct *complementary observability subspaces* at each step of our algorithm. When $m+p\leqq n$ the property of complementary observability must be released. The fact that some eigenvalues might be induced by constructing noncomplementary observability subspaces must be viewed as a positive fact as far as these induced eigenvalues can be almost arbitrarily chosen. While some of the results obtained in this paper are sufficiently general so as to cope with these problems, some others are to be amended for instance by introducing degeneracy conditions on systems. In this field, numerous problems are still open (see Magni and Champetier [11] for further discussions).

REFERENCES

[1] B. D. O. ANDERSON, *A note on transmission zeros of a transfer function matrix*, IEEE Trans. Automat. Control, 25 (1980), pp. 589-591.

[2] E. J. DAVISON AND S. H. WANG, *On pole assignment in linear multivariable systems using output feedback*, IEEE Trans. Automat. Control, 20 (1975), pp. 516-518.

[3] L. R. Fletcher, *Placement des valeurs propres pour les systèmes linéaires multivariables*, C.R. Acad. Sci. Paris, Ser. A, 289 (1979), pp. 499-501.

[4] ———, *An intermediate algorithm for pole placement by output feedback in linear multivariable control systems*, Internat. J. Control, 31 (1980), pp. 1121-1136.

[5] ——— *On exact pole assignment by output feedback* 2, Internat. J. Control, 45 (1987), pp. 2009-2019.

[6] L. R. FLETCHER AND J. F. MAGNI, *On exact pole assignment by output feedback* 1, Internat. J. Control, 45 (1987), pp. 1995-2007.

[7] H. KIMURA, *Pole assignment by gain output feedback*, IEEE Trans. Automat. Control, 20 (1975), pp. 509-516.

[8] J. F. MAGNI, *A necessary condition for pole assignment by output feedback*, in Proc. 24th IEEE Conference on Decision and Control, Fort Lauderdale, FL, December 1985, IEEE Computer Society, Washington, DC, pp. 194-195.

[9] ———, *Placement de pôles par retour de sortie, une condition nécessaire et suffisante*, Tech. Report 133/85RA, ERT, BP 40L5, F31055, Toulouse, France, September 1985.

[10] ———, *On exact pole assignment by output feedback* 3, Internat. J. Control, 45 (1987), pp. 2021-2033.

[11] J. F. MAGNI AND C. CHAMPETIER, *A general framework for pole assignment algorithms*, in Proc. 27th IEEE Conference on Decision and Control, Austin, December 1988; IEEE Trans. Automat. Control., to appear.

[12] R. R. MIELKE AND S. R. LIBERTY, *An eigenvalue/eigenvector assignment algorithm using output feedback*, in Proc. IEEE Southeastern Conference, Orlando, FL, IEEE Computer Society, Washington, DC, 1963, pp. 577-581.

[13] B. C. MOORE, *On the flexibility offered by state feedback in multivariable system beyond closed loop eigenvalue assignment*, IEEE Trans. Automat. Control, 21 (1976), pp. 659-692.

[14] J. M. SCHUMACHER, *Compensator synthesis using* (C, A, B)*-pairs*, IEEE Trans. Automat. Control, 25 (1980), pp. 1133-1138.

[15] ———, *Algebraic characterizations of almost invariance*, Internat. J. Control, 38 (1983), pp. 107-124.

[16] J. C. WILLEMS AND C. COMMAULT, *Disturbance decoupling by measurement feedback with stability or pole placement*, SIAM J. Control Optim., 19 (1979), pp. 490-504.

[17] W. M. WONHAM, *Linear Multivariable Control, a Geometric Approach*, Springer-Verlag, New York, Heidelberg, Berlin, 1979.

[18] YU-FAN ZHENG AND ZHENG-ZHI HAN, *What can be done for linear systems by feedback*, in Proc. 9th World IFAC Conference, Budapest, Hungary, V:1-6, 1984.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 866-877, July 1991

STRUCTURE AND CLASSIFICATION THEOREMS OF FINITE-DIMENSIONAL EXACT ESTIMATION ALGEBRAS*

RUI-TAO DONG†, LUEN-FAI TAM‡, WING SHING WONG§, AND STEPHEN S.-T. YAU¶

Abstract. Estimation algebra turns out to be a crucial concept in the investigation of finite-dimensional nonlinear filters. In an earlier paper by the authors, a necessary and sufficient algebraic condition was derived for an exact estimation algebra to be finite-dimensional. In this paper, the investigation of the properties of finite-dimensional exact estimation algebras is continued, and some structure and partial classification theorems for such algebras are proved.

Key words. nonlinear filters, solvable Lie algebra, estimation algebra, elliptic partial differential equation

AMS(MOS) subject classifications. 17B30, 35J15, 60G35, 93E11

1. Introduction. In a previous paper [1], we introduced the concept of an exact estimation algebra. A simple algebraic necessary and sufficient condition was proved for an exact estimation algebra to be finite-dimensional. We also provided a detailed examination of the relationship between finite-dimensional exact estimation algebras and finite-dimensional nonlinear filters. This paper is in essence a continuation of our earlier study of exact estimation algebra, and we strongly recommend that the readers familiarize themselves with the results in [1]. However, every effort will be made to make this paper as self-contained as possible without too much duplication of the previous paper.

In this paper, we will prove some structure and partial classification theorems of exact finite-dimensional estimation algebras. The class of nonlinear filtering systems with an exact estimation algebra can be characterized by the solutions of some family of Riccati partial differential equations. These equations are the focal point of this study. We will provide two alternative existence proofs of these equations and examine their uniqueness properties.

2. Basic concepts. In this section, we will recall some basic concepts and results from [1]. The idea of using estimation algebras to construct finite-dimensional nonlinear filters was first proposed in Brockett and Clark [2], Brockett [3], and Mitter [4].[1] The motivation came from the Wei–Norman approach [5] of using Lie algebraic ideas to solve linear time-varying differential equations.

Consider a filtering problem based on the following signal observation model:

$$\begin{aligned} dx(t) &= f(x(t))\,dt + g(x(t))\,dv(t), \qquad x(0) = x_0, \\ dy(t) &= h(x(t))\,dt + dw(t), \qquad y(0) = 0, \end{aligned} \tag{2.0}$$

in which x, v, y, and w, are respectively, $\mathbb{R}^n$, $\mathbb{R}^p$, $\mathbb{R}^m$, and $\mathbb{R}^m$-valued processes, and v and w have components that are independent, standard Brownian processes. We further

* Received by the editors April 10, 1989; accepted for publication (in revised form) May 24, 1990. This work was partially supported by U.S. Army Research Office grant DAAL03-89-K-0123.

† Mathematical Sciences Research Institute, 1000 Centennial Drive, Berkeley, California 94720.

‡ Department of Mathematics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong.

§ Room 3M-329, AT&T Bell Laboratories, Holmdel, New Jersey 07733.

¶ Department of Mathematics, University of Illinois at Chicago, Box 4348, Chicago, Illinois 60680.

[1] This reference was inadvertently omitted in [1].

assume that $n=p$, f, h are C^∞ smooth, and that g is an orthogonal matrix. We will refer to $x(t)$ as the state of the system at time t and to $y(t)$ as the observation at time t.

Let $\rho(t,x)$ denote the conditional probability density of the state given the observation $\{y(s): 0\leqq s\leqq t\}$. It is well known (see [6], for example) that $\rho(t,x)$ is given by normalizing a function $\sigma(t,x)$, which satisfies the following Duncan-Mortensen-Zakai equation:

$$d\sigma(t,x)=L_0\sigma(t,x)\,dt+\sum_{i=1}^m L_i\sigma(t,x)\,dy_i(t),\qquad \sigma(0,x)=\sigma_0, \tag{2.1}$$

where

$$L_0=\frac{1}{2}\sum_{i=1}^n\frac{\partial^2}{\partial x_i^2}-\sum_{i=1}^n f_i\frac{\partial}{\partial x_i}-\sum_{i=1}^n\frac{\partial f_i}{\partial x_i}-\frac{1}{2}\sum_{i=1}^m h_i^2$$

and for $i=1,\cdots,m$, L_i is the zero-degree differential operator of multiplication by h_i.[2] σ_0 is the probability density of the initial point x_0. In this paper, we will assume that σ_0 is a C^∞ function.

Equation (2.1) is a stochastic partial differential equation. In real applications, we are interested in constructing state estimators from observed sample paths with some property of robustness. In [7] Davis studied this problem and proposed some robust algorithms. In our case, his basic idea reduces to defining a new unnormalized density

$$\xi(t,x)=\exp\left(-\sum_{i=1}^m h(x)_i y_i(t)\right)\sigma(t,x).$$

It is easy to show that $\xi(t,x)$ satisfies the following time-varying partial differential equation

$$\begin{aligned}\frac{d\xi(t,x)}{dt}&=L_0\xi(t,x)+\sum_{i=1}^m y_i(t)[L_0,L_i]\xi(t,x)\\&\quad+\frac{1}{2}\sum_{i=1}^m y_i^2(t)[[L_0,L_i],L_i]\xi(t,x),\qquad \xi(0,x)=\sigma_0,\end{aligned}\tag{2.2}$$

where $[\cdot,\cdot]$ is the Lie bracket defined as follows.

Definition. If X and Y are differential operators, the Lie bracket of X and Y, $[X,Y]$, is defined by

$$[X,Y]\zeta=X(Y\zeta)-Y(X\zeta)$$

for any C^∞ function ζ.

Definition. The estimation algebra $\mathbf{E}$ of a filtering problem (2.0), is defined to be the Lie algebra generated by $\{L_0,L_1,\cdots,L_m\}$, or $\mathbf{E}=\langle L_0,L_1,\cdots,L_m\rangle_{L.A.}$. If in addition there exists a potential function ϕ such that $f_i=\partial\phi/\partial x_i$, for all $1\leqq i\leqq n$, then the estimation algebra is called exact.

From now on, unless stated otherwise, we assume the estimation algebra of (2.0) is exact. We use ∇p to denote the column vector $(\partial p/\partial x_1,\cdots,\partial p/\partial x_n)^T$. Hence, $\nabla\phi=f$.

[2] If p is a vector, we use the notation p_i to represent the ith component of p.

Define $D_i = \partial/\partial x_i - f_i$, and $\eta = \sum_{i=1}^n \partial f_i/\partial x_i + \sum_{i=1}^n f_i^2 + \sum_{i=1}^m h_i^2$. Then,

$$L_0 = \frac{1}{2}\left(\sum_{i=1}^n D_i^2 - \eta\right). \tag{2.3}$$

Recall that $f_i = \partial\phi/\partial x_i$. Hence,

$$\eta = \Delta\phi + |\nabla\phi|^2 + \sum_{i=1}^m h_i^2. \tag{2.4}$$

In [8] the two matrices Ω and J_η were introduced. Ω is the matrix whose i, j element is $\partial f_j/\partial x_i - \partial f_i/\partial x_j$. Note that all exact estimation algebras are characterized by the fact that $\Omega = 0$. $J_\eta = [\partial^2\eta/\partial x_i\, \partial x_j]$ is the Hessian matrix of η.

In [1], we proved the following structure theorems.

THEOREM 1. *Let* **E** *be a finite-dimensional exact estimation algebra. Then* $h_1, \cdots, h_m$ *are polynomials of degree at most one.*

THEOREM 2. *Let* $F(x_1, \cdots, x_n)$ *be a* C^∞ *function on* $\mathbb{R}^n$. *Suppose that there exists a path* $C: \mathbb{R} \to \mathbb{R}^n$ *and* $\delta > 0$ *such that* $\lim_{t\to\infty} \|C(t)\| = \infty$ *and* $\lim_{t\to\infty} \sup_{B_\delta(C(t))} F = -\infty$, *where* $B_\delta(C(t)) = \{x \in \mathbb{R}^n \mid \|x - C(t)\| < \delta\}$. *Then there is no* C^∞ *function* ψ *on* $\mathbb{R}^n$ *satisfying*

$$\Delta\psi + |\nabla\psi|^2 = F.$$

COROLLARY 1. *Let* $F(x_1, \cdots, x_n)$ *be a polynomial on* $\mathbb{R}^n$. *Suppose that there exists a polynomial path* $C: \mathbb{R} \to \mathbb{R}^n$ *such that* $\lim_{t\to\infty} \|C(t)\| = \infty$ *and* $\lim_{t\to\infty} F \circ C(t) = -\infty$. *Then there is no* C^∞ *function* ψ *on* $\mathbb{R}^n$ *satisfying*

$$\Delta\psi + |\nabla\psi|^2 = F. \tag{2.5}$$

THEOREM 3. *Suppose* **E** *is an exact estimation algebra. Then,* **E** *is finite-dimensional if and only if* $\nabla h_i^T J_\eta^j$ *is a constant for* $1 \leqq i \leqq m$ *and all* $j = 0, 1, \cdots$.

THEOREM 4. *Suppose* **E** *is an exact finite-dimensional estimation algebra. Then it has a basis consisting of one second-degree differential operator* L_0, *first-degree differential operator(s) with constant coefficients for the* $\partial/\partial x_i$ *terms,*[3] *and zero-degree differential operator(s) affine in x. Moreover, if X and Y are in* **E** *with degree less than or equal to one, then* $[X, Y]$ *is a constant.*

Theorem 5 follows from Theorem 4.

THEOREM 5. *An exact finite-dimensional estimation algebra is solvable.*

To show the relevancy of studying finite-dimensional exact estimation algebra in nonlinear filtering problems, we proved in [1] that a system defined by (2.0) with a finite-dimensional exact estimation algebra admits a universal finite-dimensional filter and provided an explicit Lie-algebraic method to construct such a filter.

Given the importance of the estimation algebra, a natural question arises as to whether we can classify all finite-dimensional exact estimation algebras up to Lie-algebraic isomorphism. Theorems 4 and 5 provide a starting point for solving this problem. In Theorem 6, we provide a more explicit structure theorem for an important subclass of finite-dimensional exact estimation algebras. A second question that arises naturally is whether we can classify all filtering systems with finite-dimensional exact estimation algebras up to state-space diffeomorphism. This is apparently a very difficult problem and requires a careful study of partial differential equations of the type (2.4).

[3] This clarifies the original statement of Theorem 5 of [1].

The connection between these types of equations and the nonlinear filtering problem was first noted by Benes (see [9]). The properties of these equations, however, are not well known. In Theorems 9 and 12, we provide some answers in regard to the existence and uniqueness of the solutions of these types of equations. Our result here is far from providing a reasonable classification theory of systems with finite-dimensional exact estimation algebras, but it may be viewed as a necessary first step.

3. Classification theorems. If **E** is finite-dimensional, then the matrix

$$M=[\nabla h_1,\cdots,\nabla h_m,J_\eta\nabla h_1,\cdots,J_\eta\nabla h_m,J_\eta^2\nabla h_1,\cdots,J_\eta^2\nabla h_m,\cdots]\tag{3.0}$$

is a constant matrix and

$$h_1,\cdots,h_m,\nabla h_1^T\nabla\eta,\cdots,\nabla h_m^T\nabla\eta,\nabla h_1^TJ_\eta\nabla\eta,\cdots,\nabla h_m^TJ_\eta\nabla\eta,\cdots$$

are all linear functions in **E**. If the rank of M is n, we say that the corresponding estimation algebra has full rank. In this case, it is easy to describe the Lie algebra structure of the estimation algebra.

THEOREM 6. *Suppose* **E** *is of maximal rank. Then it is a real vector space of dimension* $2n+2$ *with basis given by* 1, x_1, $x_2,\cdots,x_n$, $D_1,\cdots,D_n$, *and* L_0. *Moreover,* η *is a polynomial of degree at most two and the quadratic part of* $\eta-\sum_{i=1}^m h_i^2$ *is positive semidefinite.*

Proof. Since the columns of M represent gradient vectors of functions in **E** and M is a constant matrix with rank n, there are constants c_i's such that x_i+c_i is in **E** for $i=1,\cdots,m$. It is easy to show the following relations:

$$[L_0,x_i+c_i]=\frac{1}{2}\left[\sum_{j=1}^n D_j^2,x_i+c_i\right]=D_i,$$

$$[D_i,x_j+c_j]=\begin{cases}1 & \text{if } i=j,\\ 0 & \text{if } i\neq j,\end{cases}$$

$$[L_0,D_i]=\frac{1}{2}\left[\sum_{j=1}^n D_j^2,D_i\right]+\frac{1}{2}[D_i,\eta]=\frac{1}{2}\frac{\partial\eta}{\partial x_i}.$$

$\partial\eta/\partial x_i$ is a polynomial of degree at most one, for all $1\leqq i\leqq n$. Hence **E** is a real vector space spanned by 1, $x_1,\cdots,x_n$, $D_1,\cdots,D_n$ and L_0. The fact that the quadratic part of $\eta-\sum_{i=1}^m h_i^2$ is positive semidefinite again follows from Theorem 2. □

Theorem 6 implies that all exact finite-dimensional estimation algebras with maximal rank come from Benes filters (see [9] for details concerning Benes filters).

For any filtering system defined in (2.0) with an exact estimation algebra, (2.4) assigns a characteristic η. Theorem 6 implies that if the estimation algebra is finite-dimensional with maximal rank, then this mapping maps the given system to a quadratic polynomial. In order to develop a classification of systems with finite-dimensional estimation algebras, we need to know the range of this mapping restricted to such systems. We also need to understand the properties of the inverse of this mapping. In the following, we will provide some partial results to these questions. The key to these questions is a complete understanding of the existence and uniqueness properties of (2.5).

Let q be a C^∞ function defined on $\mathbb{R}^n$. Extend $-\Delta+q$ in the standard way to act on a closed subspace of $L^2(\mathbb{R}^n)$. It follows from the definition that the first eigenvalue λ_1 of the operator $-\Delta+q$ is equal to

$$\lambda_1=\inf_\phi\frac{\int|\nabla\phi|^2\,dx+\int q\phi^2\,dx}{\int\phi^2\,dx},\tag{3.1}$$

where infimum is taken on all nonzero C^∞ functions with compact support. The following theorem by Fischer-Colbrie and Schoen is well known (see [10]).

THEOREM 7 [10]. *Let q be a C^∞ function defined on $\mathbb{R}^n$. Then there exists a positive function ζ satisfying the equation $\Delta\zeta - q\zeta = 0$ on $\mathbb{R}^n$ if and only if the first eigenvalue λ_1 of $-\Delta + q$ on $\mathbb{R}^n$ is nonnegative.*

Assume now that the estimation algebra is finite-dimensional and has maximal rank. Then by Theorem 6, we know that

$$\eta - \sum_{i=1}^{m} h_i^2 = q, \tag{3.2}$$

where q is a polynomial of degree two with quadratic part positive semidefinite. Recall that

$$\begin{aligned}\eta &= \sum_{i=1}^{m} \frac{\partial f_i}{\partial x_i} + \sum_{i=1}^{n} f_i^2 + \sum_{i=1}^{m} h_i^2 \\ &= \Delta\phi + |\nabla\phi|^2 + \sum_{i=1}^{m} h_i^2.\end{aligned}$$

Putting this into (3.2), we have

$$\Delta\phi + |\nabla\phi|^2 = q. \tag{3.3}$$

Let $u = e^\phi$. Then $\partial u/\partial x_i = (\partial\phi/\partial x_i)\, e^\phi$ and $\partial^2 u/\partial x_i^2 = (\partial^2\phi/\partial x_i^2 + (\partial\phi/\partial x_i)^2)\, e^\phi$, hence

$$\Delta u - qu = 0. \tag{3.4}$$

We observe that (3.3) has a C^∞-solution ϕ if and only if (3.4) has a C^∞ positive solution u.

THEOREM 8. *Let q be a quadratic polynomial in $x_1, \cdots, x_n$. Let λ_1 be the first eigenvalue of the operator $-\Delta + q$. Then λ_1 is nonnegative if and only if under an orthogonal transformation and a translation, q can be written in the form*

$$\sum_{i=1}^{n} a_i x_i^2 - c,$$

where a_i and c are constants, $a_i \geqq 0$, and $c \leqq \sum_{i=1}^{n} \sqrt{a_i}$.

Proof. Suppose that $x = (x_1, \cdots, x_n)^T = Ay - y_0$, where A is an orthogonal matrix and y_0 is a constant vector. Then $\Delta_y = \Delta_x$, and the first eigenvalue of the operator $-\Delta_x + q$ is nonnegative if and only if the first eigenvalue of $-\Delta_y + \tilde{\eta}$ is nonnegative where $\tilde{\eta}(y) = \eta(x(y))$. Hence after an orthogonal transformation and a translation, we may assume that

$$q(x) = \sum_{i=1}^{l} a_i x_i^2 + \sum_{i=l+1}^{n} b_i x_i - c,$$

where a_i, b_i, and c are constants, $a_i \neq 0$, for $i = 1, \cdots, l$.

By Theorem 7, we know that $\lambda_1 \geqq 0$ if and only if (3.4) has C^∞ positive solution if and only if (3.3) has C^∞ solution. In view of Theorem 2, this implies that $b_i = 0$ and we have that $a_i \geqq 0$ for $i = 1, \cdots, n$. Hence it remains to prove that the first eigenvalue of the operator $-\Delta + r - c$ is nonnegative if and only if $c \leqq \sum_{i=1}^{n} \sqrt{a_i}$, where $r = \sum_{i=1}^{n} a_i x_i^2$. This is equivalent to proving that the first eigenvalue λ_1' of the operator $-\Delta + r$ is $\sum_{i=1}^{n} \sqrt{a_i}$.

Denote $c_0 = \sum_{i=1}^n \sqrt{a_i}$. Let ξ be a C^∞ function with compact support. Then

$$\begin{aligned}
\int |\nabla \xi|^2\, dx + \int r\xi^2\, dx &= \int \sum_{i=1}^n \left(\left(\frac{\partial \xi}{\partial x_i} \right)^2 + a_i x_i^2 \xi^2 \right) dx \\
&= \int \sum_{i=1}^n \left(\frac{\partial \xi}{\partial x_i} + \sqrt{a_i}\, x_i \xi \right)^2 dx - 2 \int \left(\sum_{i=1}^n \sqrt{a_i}\, x_i \xi \frac{\partial \xi}{\partial x_i} \right) dx \\
&= \int \sum_{i=1}^n \left(\frac{\partial \xi}{\partial x_i} + \sqrt{a_i}\, x_i \xi \right)^2 dx \\
&\qquad - \int \sum_{i=1}^n \frac{\partial}{\partial x_i} (\sqrt{a_i}\, x_i \xi^2)\, dx + \int \left(\sum_{i=1}^n \sqrt{a_i} \right) \xi^2\, dx \\
&= \int \sum_{i=1}^n \left(\frac{\partial \xi}{\partial x_i} + \sqrt{a_i}\, x_i \xi \right)^2 dx + c_0 \int \xi^2\, dx \\
&\geqq c_0 \int \xi^2\, dx.
\end{aligned}$$

Hence $\lambda_1' \geqq c_0$. On the other hand,

$$\chi(x) = \prod_{i=1}^n \exp\left(-\frac{\sqrt{a_i}\, x_i^2}{2} \right)$$

is an eigenfunction of $-\Delta + r$ with eigenvalue c_0, so $c_0 \geqq \lambda_1'$. Hence $c_0 = \lambda_1'$. □

THEOREM 9. *Suppose* $\mathbf{E}$ *is a finite estimation algebra of maximal rank. Then under an orthogonal transformation and a translation,* η *can be written in the form*

$$\sum_{i=1}^m h_i^2 + \sum_{i=1}^n a_i x_i^2 - c,$$

where a_i *and* c *are constants,* $a_i \geqq 0$, *and* $c \leqq \sum_{i=1}^n \sqrt{a_i}$.

Proof. This result follows from Theorems 7 and 8. □

4. Alternative proof. Theorem 9 provides a constraint that the coefficients of (2.0) must satisfy so that the system has a finite-dimensional estimation algebra of maximal rank. It is a first step in providing some classification results of all finite-dimensional exact estimation algebras. In the following, we provide an alternative proof of these results by applying a technique pioneered by Li and Yau [11]. In fact, Theorem 12 sharpens the results stated in Theorem 9.

THEOREM 10. *Consider the following equation*:

$$\Delta \xi + |\nabla \xi|^2 = \sum_{ij=1}^n a_{ij} x_i x_j - c, \tag{4.0}$$

where $(x_1, \cdots, x_n) \in \mathbb{R}^n$, $c \in \mathbb{R}$ *and the constant matrix* $A = (a_{ij})$ *is positive semidefinite. Then for any smooth solution* ξ *of* (4.0) *defined on* $\mathbb{R}^n$, $|\nabla \xi|$ *has at most linear growth, namely,*

$$|\nabla \xi(x)| \leqq C(1 + |x|), \qquad x \in \mathbb{R}^n$$

for some constant C.

Proof. Let $u = -\xi$. After orthogonal change of coordinates, (4.0) becomes

$$-\Delta u + |\nabla u|^2 = \sum_{i=1}^n \lambda_i x_i^2 - c. \tag{4.1}$$

Let $v=\sum_{i=1}^n \frac{1}{2}\sqrt{\lambda_i}\, x_i^2$. It is easy to see that

$$-\Delta v+|\nabla v|^2=\sum_{i=1}^n \lambda_i x_i^2-c_0, \tag{4.2}$$

where $c_0=\sum_{i=1}^n \sqrt{\lambda_i}$. Let $w=u-v$. Then subtracting (4.2) from (4.1), we get

$$\begin{aligned} c_0-c&=-\Delta u+\Delta v+|\nabla u|^2-|\nabla v|^2\\ &=-\Delta(u-v)+\nabla(u-v)\cdot\nabla(u-v)+2\nabla(u-v)\cdot\nabla v\\ &=-\Delta w+|\nabla w|^2+2\nabla w\cdot\nabla v, \end{aligned} \tag{4.3}$$

where $x\cdot y$ represents the standard inner product between vectors x and y. Denote $F=|\nabla w|^2$. Direct computation yields

$$\begin{aligned} \Delta F&=\Delta(\nabla w\cdot\nabla w)\\ &=2\sum_{i,j=1}^n\left(\frac{\partial^2 w}{\partial x_i\,\partial x_j}\right)^2+2(\nabla\Delta w)\cdot\nabla w\\ &\geqq\frac{2}{n}|\Delta w|^2+2\nabla(F+2\nabla v\cdot\nabla w+c-c_0)\cdot\nabla w. \end{aligned} \tag{4.4}$$

If $\nabla^2 v$ and $\nabla^2 w$ represent the Hessian of v and w, respectively, then

$$\begin{aligned} 4\nabla(\nabla v\cdot\nabla w)\cdot\nabla w&=4[\nabla^2 v\nabla w+\nabla^2 w\nabla v]\cdot\nabla w\\ &\geqq 4[\nabla^2 w\nabla v]\cdot\nabla w=4[\nabla^2 w\nabla w]\cdot\nabla v\\ &=2\nabla(\nabla w\cdot\nabla w)\cdot\nabla v=2\nabla F\cdot\nabla v. \end{aligned} \tag{4.5}$$

Putting (4.5) into (4.4), we have

$$\begin{aligned} \Delta F&\geqq\frac{2}{n}(F+2\nabla v\cdot\nabla w+c-c_0)^2+2\nabla F\cdot\nabla w+2\nabla F\cdot\nabla v\\ &\geqq\frac{2}{n}F^2+\frac{4}{n}F(2\nabla v\cdot\nabla w+c-c_0)+2\nabla F\cdot\nabla(v+w)\\ &\geqq\frac{2}{n}F^2-\frac{8}{n}F^{3/2}|\nabla v|+2\nabla F\cdot\nabla(v+w)+\frac{4(c-c_0)}{n}F. \end{aligned} \tag{4.6}$$

Denote $r^2=\sum_{i=1}^n x_i^2$. For $a>0$, the function $(a^2-r^2)^2F$ achieves its maximum at $x_0\in B_a(0)=\{x\in\mathbb{R}^n:|x|<a\}$. At that point,

$$\nabla[(a^2-r^2)^2F]=0,$$

which implies

$$4rF\nabla r=(a^2-r^2)\nabla F. \tag{4.7}$$

Also at the point x_0,

$$\begin{aligned} 0&\geqq\Delta[(a^2-r^2)^2F]\\ &=(a^2-r^2)^2\Delta F+2\nabla(a^2-r^2)^2\cdot\nabla F+F\Delta(a^2-r^2)^2\\ &=(a^2-r^2)^2\Delta F-8(a^2-r^2)r\nabla r\cdot\nabla F+[8r^2-4n(a^2-r^2)]F. \end{aligned} \tag{4.8}$$

Using (4.7), we get

$$(a^2-r^2)^2\Delta F-24r^2F-4n(a^2-r^2)F\leqq 0. \tag{4.9}$$

Using (4.6), we have

$$\begin{aligned}(a^2-r^2)^2\left[\frac{2}{n}F^2-\frac{8}{n}F^{3/2}|\nabla v|+2\nabla F\cdot\nabla(v+w)+\frac{4(c-c_0)}{n}F\right]&\\ -[24r^2+4n(a^2-r^2)]F\leqq 0.&\end{aligned} \tag{4.10}$$

Dotting (4.7) with $\nabla(v+w)$, we get

$$\begin{aligned}(a^2-r^2)\nabla(v+w)\cdot\nabla F&=4rF\nabla r\cdot(\nabla v+\nabla w)\\ &\geqq -4rF|\nabla v|-4rF^{3/2}.\end{aligned} \tag{4.11}$$

Putting (4.11) back into (4.10) and dividing it by F, we have

$$\begin{aligned}\frac{2}{n}(a^2-r^2)^2F-\frac{8}{n}(a^2-r^2)^2F^{1/2}|\nabla v|-(a^2-r^2)[8r|\nabla v|+8rF^{1/2}]&\\ +4(a^2-r^2)^2\frac{c-c_0}{n}-[24r^2+4n(a^2-r^2)]\leqq 0.&\end{aligned} \tag{4.12}$$

By denoting $M=(a^2-r^2)F^{1/2}$, (4.12) becomes

$$\begin{aligned}&\frac{2}{n}M^2-\left[\frac{8}{n}(a^2-r^2)|\nabla v|+8r\right]M\\ &\quad+\left[4(a^2-r^2)^2\frac{c-c_0}{n}-4n(a^2-r^2)-8r(a^2-r^2)|\nabla v|-24r^2\right]\leqq 0.\end{aligned}$$

Noting the fact that $|\nabla v|\leqq c_1r$ and $r\leqq a$, we can see that

$$M\leqq c_2a^3,$$

where c_1 and c_2 are constants. The inequality

$$\begin{aligned}M&=\max_{|x|\leqq a}(a^2-r^2(x))F^{1/2}(x)\\ &\geqq\max_{|x|\leqq a/2}(a^2-|x|^2)F^{1/2}(x)\\ &\geqq\max_{|x|\leqq a/2}\left[a^2-\left(\frac{a}{2}\right)^2\right]F^{1/2}(x)\\ &=\frac{3}{4}a^2\max_{|x|\leqq a/2}|\nabla w|\end{aligned}$$

yields the estimate

$$\max_{|x|\leqq a/2}|\nabla w|\leqq C_3a. \tag{4.13}$$

Combining (4.13) with the relation $w=u-v$, we can conclude that $|\nabla u|$ has at most linear growth. □

Remark. We can also deduce the above theorem by making use of Theorem 1.3 of [11].

THEOREM 11. *If $c<\sqrt{\lambda}$, and $\lambda>0$, then*

$$-u''+(u')^2=\lambda x^2-c, \quad u(0)=a, \quad u'(0)=b, \tag{4.14}$$

has a global solution for any a and small $|b|$.

Proof. Let $v=u'$. We have

$$v'=v^2-\lambda x^2+c, \qquad v(0)=b.$$

Suppose (A, B) is the maximum open interval containing zero, such that v exists. Define two auxiliary functions:

$$v_+(x)=\varepsilon x+k, \qquad v_-(x)=-\delta x-l.$$

We have that

$$\begin{aligned} v'_+-v_+^2+\lambda x^2-c&=\varepsilon-(\varepsilon x+k)^2+\lambda x^2-c\\ &=(\lambda-\varepsilon^2)x^2-2k\varepsilon x+(\varepsilon-c-k^2). \end{aligned}$$

Choose $\varepsilon>0$, such that

$$\lambda-\varepsilon^2>0 \quad\text{and}\quad \varepsilon-c>0.$$

This is possible because $\sqrt{\lambda}>c$. Choose $k>0$ small enough, so that

$$(\lambda-\varepsilon^2)x^2-2k\varepsilon x+(\varepsilon-c-k^2)>0, \qquad x\in[0, B].$$

By the standard comparison theorem [12], we have

$$v(x)<v_+(x) \quad \text{for } x\in[0, B),$$

as long as $v(0)=b<k=v_+(0)$.

Similarly, we can show that if δ is sufficiently large, so that

$$\lambda-\delta^2<0 \quad\text{and}\quad \delta+c>0,$$

then

$$v'_--v_-^2+\lambda x^2-c=(\lambda-\delta^2)x^2-2l\delta x-(\delta+c+l^2)<0$$

for all $x\in[0, B]$ and $l\geqq 0$.

The comparison theorem again implies that

$$v(x)>v_-(x) \quad \text{for } x\in[0, B),$$

if $v(0)=b>-l=v_-(0)$.

This implies that when $-l<b<k$, $B=\infty$. Otherwise, as $v(B)$ is bounded, we can extend v beyond B, a contradiction to the hypothesis that (A, B) is the maximal interval on which v is defined.

Similar arguments show that $A=-\infty$ when $|b|$ is sufficiently small. □

Remark. If $\lambda=0$, then we can prove by direct integration that there is a global solution to (4.14) if $|b|$ is small enough.

THEOREM 12. *Consider the following equation:*

$$\Delta\xi+|\nabla\xi|^2=\sum_{i,j=1}^{n} a_{ij}x_ix_j-c, \tag{4.15}$$

where $(x_1, \cdots, x_n)\in\mathbb{R}^n$, $c\in\mathbb{R}$ and the constant matrix $A=(a_{ij})$ is positive semidefinite. Let $\{\lambda_1, \cdots, \lambda_n\}$ be the eigenvalues of A and $c_0=\sum_{i=1}^n\sqrt{\lambda_i}$. Then we have the following:

(I) (*Existence*). *When* $c < c_0$, *there is a family of* C^∞ *solution of* (4.15) *with* $2n$ *parameters such that* $|\nabla \xi|$ *has at most linear growth at* ∞, *namely,*

$$|\nabla \xi(x)| \leqq C(1+|x|),$$

for some constant C.

(II) (*Uniqueness*). *When* $c = c_0$, *there is a quadratic polynomial, uniquely determined up to a constant, which satisfies* (4.15). *Moreover, this is the unique solution up to a constant if either one of the following conditions holds*:

(i) rank $A = 0$ (*namely,* $A = 0$), *or*

(ii) rank $A \geqq n-2$.

(III) (*Nonexistence*). *When* $c > c_0$, *there is no smooth solution to* (4.15).

Proof. Let $u = -\xi$. After an orthogonal change of coordinates, (4.15) becomes

$$-\Delta u + |\nabla u|^2 = \sum_{i=1}^{n} \lambda_i x_i^2 - c. \tag{4.16}$$

For part (I), let $c = \sum_{i=1}^n c_i$ with $c_i < \sqrt{\lambda_i}$. By Theorem 11, there is a 2-parameter family of solution of the

$$-u_i'' + (u_i')^2 = \lambda_i x_i^2 - c_i. \tag{4.17}$$

It is easy to see that $u(x) = \sum_{i=1}^n u_i(x_i)$ satisfies (4.15) and that $|\nabla u|$ has at most linear growth at ∞. To see that there is a $2n$-parameter family of such solutions to (4.16), note that $n-1$ parameters come from the different ways of decomposing c into c_i's so that $c = \sum_{i=1}^n c_i$, n parameters come from $u_i'(0)$, and the last parameter comes from the arbitrary constant added to the whole solution.

For part (II), it is clear that there exists a uniquely determined quadratic polynomial solution. If in addition the first rank condition is satisfied, we need only to prove that the only solutions of

$$-\Delta u + |\nabla u|^2 = 0 \tag{4.18}$$

are constants. Taking $\Phi = e^{-u}$, (4.18) can be written as

$$\Delta \Phi = 0. \tag{4.19}$$

It is equivalent to prove that (4.19) has no positive solutions other than constants. However, this is well known to be the case.

Next, assume the second rank condition is satisfied, that is, rank $A \geqq n-2$. Let $v(x) = \sum_{i=1}^n \frac{1}{2}\sqrt{\lambda_i}\, x_i^2$. Note that $v(x)$ satisfies

$$-\Delta v + |\nabla v|^2 = \sum_{i=1}^{n} \lambda_i x_i^2 - c_0. \tag{4.20}$$

Subtracting (4.20) from (4.16) and letting $w = u - v$, we get

$$-\Delta w + 2\nabla v \cdot \nabla w + |\nabla w|^2 = 0. \tag{4.21}$$

Define $B(r) = \{x : |x_i| \leqq r, i = 1, 2, \cdots, n\}$. Multiplying by e^{-2v} and integrating on both sides of (4.21), we get

$$\begin{aligned} 0 &= \int_{B(r)} e^{-2v}(-\Delta w + 2\nabla v \cdot \nabla w) + \int_{B(r)} e^{-2v}|\nabla w|^2 \\ &= \int_{B(r)} -\nabla \cdot (e^{-2v}\nabla w) + \int_{B(r)} e^{-2v}|\nabla w|^2 \\ &= -\int_{\partial B(r)} e^{-2v}\nabla w \cdot \vec{n} + \int_{B(r)} e^{-2v}|\nabla w|^2, \end{aligned}$$

where $\vec{n}$ is the unit outward normal of $\partial B(r)$. By the Schwartz inequality, we get

$$\int_{B(r)} e^{-2v}|\nabla w|^2 = \int_{\partial B(r)} e^{-2v}\nabla w \cdot \vec{n} \leqq \int_{\partial B(r)} e^{-2v}|\nabla w| \leqq \sqrt{\left(\int_{\partial B(r)} e^{-2v}|\nabla w|^2\right)\left(\int_{\partial B(r)} e^{-2v}\right)}. \tag{4.22}$$

Denoting $f(r) = \int_{B(r)} e^{-2v}|\nabla w|^2$, $g(r) = \int_{B(r)} e^{-2v}$, (4.22) becomes

$$(f(r))^2 \leqq f'(r) \cdot g'(r). \tag{4.23}$$

Supposing $f(r_0) > 0$ for some r_0, we have

$$\frac{f'}{f^2} \geqq \frac{1}{g'}. \tag{4.24}$$

Integrating (4.24) over $(r_0, +\infty)$, we have

$$\infty > \frac{1}{f(r_0)} - \frac{1}{f(\infty)} = \int_{r_0}^{\infty} \frac{f'}{f^2}\, dr \geqq \int_{r_0}^{\infty} \frac{1}{g'}\, dr. \tag{4.25}$$

Other the other hand,

$$g(r) = \int_{B(r)} e^{-2v} = \prod_{i=1}^{n} \left(\int_{-r}^{r} \exp\left(-\sqrt{\lambda_i}\, x_i^2\right) dx_i\right).$$

It is easy to see that

(i) If rank $A = n$, $g'(r) \to 0$ as $r \to \infty$.

(ii) If rank $A = n - 1$, $g'(r) \to c > 0$ as $r \to \infty$.

(iii) If rank $A = n - 2$, $g'(r) \approx cr$ as $r \to \infty$.

In all of the above three cases, the right-hand side of (4.25) is divergent. The contradiction says that $f(r) \equiv 0$, namely, w is a constant. So $u = v + \text{const}$.

For part (III), the statement holds even if A is degenerate. Since $c_0 = \sum_{i=1}^{n} \sqrt{\lambda_i} < c$, we can find δ small such that $\sum_{i=1}^{n} (\sqrt{\lambda_i} + \delta) < c$. Let $v = \sum_{i=1}^{n} \frac{1}{2}(\sqrt{\lambda_i} + \delta)x_i^2$. v satisfies

$$-\Delta v + |\nabla v|^2 = \sum_{i=1}^{n} (\sqrt{\lambda_i} + \delta)^2 x_i^2 - \sum_{i=1}^{n} (\sqrt{\lambda_i} + \delta). \tag{4.26}$$

Subtracting (4.26) from (4.16) and letting $w = u - v$, we get

$$-\Delta w + 2\nabla v \cdot \nabla w \leqq 0.$$

Using the same argument as before, but without assuming the rank condition on A, we can show that $u = v + \text{const}$. So u cannot be a solution to (4.16). A contradiction and no smooth solution to (4.16) exists. □

Remark. Equation (4.15) may have other solutions in addition to those listed in part (I). Some examples are given on page 86 of [9].

REFERENCES

[1] L. Tam, W. S. Wong, and S. S.-T. Yau, *On a necessary and sufficient condition for finite dimensionality of estimation algebras*, SIAM J. Control Optim., 28 (1990), pp. 173–185.

[2] R. W. Brockett and J. M. C. Clark, *The geometry of the conditional density functions*, in Analysis and Optimization of Stochastic Systems, O. L. R. Jacobs et al., eds., Academic Press, New York, 1980, pp. 299–309.

[3] R. W. Brockett, *Nonlinear systems and nonlinear estimation theory*, in The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, 1981.

[4] S. K. MITTER, *On the analogy between mathematical problems of non-linear filtering and quantum physics*, Ricerche Automat., 10 (1979), pp. 163-216.

[5] J. WEI AND E. NORMAN, *On global representations of the solutions of linear differential equations as a product of exponentials*, Proc. Amer. Math. Soc., 15 (1964), pp. 327-334.

[6] M. H. A. DAVIS AND S. I. MARCUS, *An introduction to nonlinear filtering*, in The Mathematics of Filtering and Identification and Applications, M. Hazewinkel and J. C. Willems, eds., Reidel, Dordrecht, 1981.

[7] M. H. A. DAVIS, *On a multiplicative functional transformation arising in nonlinear filtering theory*, Z. Wahrsch. Verw. Gebiete, 54 (1980), pp. 125-139.

[8] W. S. WONG, *On a new class of finite dimensional estimation algebras*, Systems Control Lett., 9 (1987), pp. 79-83.

[9] V. BENES, *Exact finite dimensional filters for certain diffusions with nonlinear drift*, Stochastics, 5 (1981), pp. 65-92.

[10] D. FISCHER-COLBRIE AND R. SCHOEN, *The structure of complete stable minimal surfaces in* 3*-manifolds of nonnegative scalar curvature*, Comm. Pure Appl. Math., 33 (1980), pp. 199-211.

[11] P. LI AND S. T. YAU, *On the parabolic kernel of the Schrödinger operator*, Acta Math., 156 (1986), pp. 153-201.

[12] G. BIRKHOFF AND G. ROTA, *Ordinary Differential Equations*, John Wiley, New York, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 878-894, July 1991

008

STRUCTURE AT INFINITY OF STRUCTURED DESCRIPTOR SYSTEMS AND ITS APPLICATIONS*

KAZUO MUROTA† AND JACOB W. VAN DER WOUDE‡

Abstract. The generic structure at infinity of the transfer matrix of a descriptor system is investigated under the physically reasonable assumption that the coefficients in the equations are classified into independent physical parameters and dimensionless constants. For such a structured descriptor system, the generic structure at infinity is characterized in terms of an independent assignment (or weighted matroid intersection) problem. This leads to necessary and sufficient conditions for the generic solvability of the exact matching problem for a (singular) descriptor system and for the generic solvability of the disturbance decoupling problem for a nonsingular descriptor system. The obtained conditions can be checked by efficient matroid-theoretic algorithms.

Key words. structure at infinity, disturbance decoupling, model matching, descriptor system, structural approach, mixed matrix, matroid-theoretic algorithm

AMS(MOS) subject classifications. 93C05, 05C50, 15A15, 93A15, 94C15

1. Introduction. In this paper we present a structural approach to the structure at infinity of the transfer matrix of a general class of physical systems that have a so-called descriptor representation. This research is also motivated by the disturbance decoupling problem (DDP) and the exact model matching problem (EMMP), which are counted among the fundamental problems in control theory.

To introduce the DDP for the general class of descriptor systems, we first consider the well-known linear finite-dimensional time-invariant system described by

$$\dot{\mathbf{x}}(t)=A\mathbf{x}(t)+B\mathbf{u}(t)+G\mathbf{d}(t), \qquad \mathbf{y}(t)=C\mathbf{x}(t) \tag{1}$$

with state $\mathbf{x}(t)\in\mathbf{R}^n$, input $\mathbf{u}(t)\in\mathbf{R}^m$, disturbance $\mathbf{d}(t)\in\mathbf{R}^l$ and output $\mathbf{y}(t)\in\mathbf{R}^p$. The DDP for system (1) consists of finding a constant matrix K such that with the feedback $\mathbf{u}(t)=K\mathbf{x}(t)$ the transfer matrix of the closed-loop system satisfies

$$C(sI-A-BK)^{-1}G=0. \tag{2}$$

A geometric condition for the solvability of DDP is given by Wonham [31] in terms of (A, B)-invariant subspaces. Furthermore, it is shown in Emre and Hautus [5] (see also Bhattacharyya, Gomes, and Howze [3, Thm. 3]) that DDP is equivalent to EMMP, that is, that DDP is solvable if and only if there exists a strictly proper rational matrix $X(s)$ such that

$$C(sI-A)^{-1}BX(s)=C(sI-A)^{-1}G. \tag{3}$$

In the spirit of the structural or generic approach initiated by Lin [15], the generic solvability of DDP is considered by van der Woude [32], and independently by Commault, Dion, and Perez [4], under the assumption that the nonzero entries in the coefficient matrices in (1) are independent parameters. A graph-theoretic necessary and sufficient condition for the generic solvability of DDP is then derived on the basis of two observations. One is that the solvability of EMMP can be expressed in terms

* Received by the editors November 1, 1989; accepted for publication (in revised form) May 21, 1990.

† Department of Mathematical Engineering and Information Physics, Faculty of Engineering, University of Tokyo, Bunkyo-ku, Tokyo 113, Japan. This work was done at the Institut für Ökonometrie und Operations Research, Universität Bonn and was supported by the Alexander von Humboldt Foundation.

‡ Centre for Mathematics and Computer Science, Present address, Faculty of Technical Mathematics and Informatics, Delft University of Technology, P.O. Box 356, 2600 AJ Delft, the Netherlands.

of the structure at infinity of the two transfer matrices in (3), and the other is that the generic structure at infinity of a transfer matrix can be characterized in terms of linkings in a signal-flow type of graph associated with the system. Similar graph-theoretic characterizations of the generic structure at infinity are obtained by Suda, Wan, and Ueno [27], and for a restricted case by Kobayashi and Nakamizo [12]. Other investigations on the generic number of zeros are made by Reinschke [25], Söte [26], and Svaricek [28].

It has been gradually recognized, however, that the standard form (1) is not very suitable for representing the structure of a system in that the entries of the matrices in (1) are usually not independent but are algebraically related to one another. In this respect, the so-called descriptor form

$$F\dot{\mathbf{x}}(t)=A\mathbf{x}(t)+B\mathbf{u}(t)+G\mathbf{d}(t), \qquad \mathbf{y}(t)=C\mathbf{x}(t), \tag{4}$$

considered in this paper, is more suitable, where we assume throughout that $A-sF$ is nonsingular.

In this paper we adopt the physically plausible framework introduced by Murota [17]-[20], in which the notion of mixed matrices is used to obtain results concerning the generic or structural controllability and the fixed modes of physical systems. Here, we also use the notion of mixed matrices, but now we are interested in the generic structure at infinity of the transfer matrix of physical systems. Therefore, inspired by the previous works, we derive a characterization for the structure at infinity of the transfer matrix of a descriptor system and we present an algorithm by which the structure at infinity can be computed in an efficient way. In addition, we indicate how the algorithm can be used to check the solvability of two versions of DDP for descriptor systems.

Rephrasing the above, we discuss in this paper the structure at infinity and the solvability of DDP under the assumption that the entries of the coefficient matrices in the descriptor equations (4) are classified into independent physical parameters and dimensionless fixed constants. We develop a combinatorial characterization of matroid-theoretic nature for the structure at infinity, based on the characterization of dynamical degree given by Murota [17], [18] using the structural framework mentioned above. As an application of the obtained characterization we give a necessary and sufficient condition for the generic solvability of DDP for the descriptor system (4) with nonsingular F. This condition can be checked efficiently by a matroid-theoretic algorithm that is guaranteed to run in polynomial time in the size of the control system.

This paper is organized as follows. In § 2 we present some known facts about rational function matrices and mixed matrices. In § 3 we summarize some results on the solvability of DDP. In particular, the solvability of DDP is expressed in terms of the maximum degrees of minors of polynomial matrices associated with the system. In § 4 we describe a physically plausible mathematical model in which mixed matrices are used and on the basis of which a method of structural analysis is developed. In § 5 we show that the problem of determining the maximum degrees of minors of a polynomial mixed matrix is reduced to an independent assignment (or weighted matroid intersection) problem. Finally in § 6 we give an illustrative example.

2. Preliminaries.

2.1. Rational function matrix. Let $\mathbf{F}$ be a field, and denote by $\mathbf{F}[s]$ and $\mathbf{F}(s)$, respectively, the ring of polynomials and the field of rational functions in s over $\mathbf{F}$. A rational function $f(s)=p(s)/q(s)\in\mathbf{F}(s)$ with $p(s), q(s)\in\mathbf{F}[s]$ is called *proper* (respectively, *strictly proper*) if $\deg f(s)\leqq 0$ (respectively, $\deg f(s)<0$), where $\deg f(s)=\deg p(s)-\deg q(s)$ and $\deg(0)=-\infty$.

We call a matrix a *proper* (respectively, *strictly proper*) *rational matrix* if its entries are proper (respectively, strictly proper) rational functions. A square proper rational matrix is called *bicausal* if it is invertible and its inverse is a proper rational matrix. Since the proper rational functions form a Euclidean ring, any proper rational matrix can be brought into the Smith form [24], which in control literature is sometimes referred to as the *structure at infinity*. From this we see further that any rational matrix can be brought into the Smith-McMillan form at infinity, as stated below.

LEMMA 2.1. *Let $P(s)$ be a rational matrix. Then there exist bicausal matrices $U(s)$ and $V(s)$ such that*

$$P(s)=V(s)\begin{pmatrix}\Gamma(s) & O\\ O & O\end{pmatrix}U(s),$$

where

$$\Gamma(s)=\mathrm{diag}\,(s^{-t_1},\cdots,s^{-t_r}),$$

$r=\mathrm{rank}\,P(s)$, *and* t_k $(k=1,\cdots,r)$ *are integers with* $t_1\leqq\cdots\leqq t_r$. □

The integers t_k $(k=1,\cdots,r)$ are uniquely determined, and are referred to as the *orders of the zeros at infinity* of $P(s)$ when $P(s)$ is proper. In this paper we use the notation

$$\mathrm{ord}_k\,P=t_k \qquad (1\leqq k\leqq r)$$

and

$$\mathrm{ord}\,P=(t_1,\cdots,t_r).$$

An alternative characterization of ord P can be given in terms of the degrees of minors of the matrix P. Let R and C denote the row set and the column set of P, and $P[I,J]$ be the submatrix of P with row set $I\subseteq R$ and column set $J\subseteq C$. For a square rational matrix $P(s)$ we define

$$\delta(P)=\deg_s\det P(s).$$

Furthermore, for a (possibly nonsquare) rational matrix $P(s)$ and for $I_0\subseteq R$, $J_0\subseteq C$ and $k\geqq\max\,(|I_0|,|J_0|)$ we define

$$\delta_k(P;I_0,J_0)=\max\,\{\delta(P[I,J])\,|\,I\supseteq I_0,J\supseteq J_0,|I|=|J|=k\} \tag{5}$$

and

$$\delta_k(P)=\delta_k(P,\emptyset,\emptyset).$$

Using the Cauchy-Binet formula, the following can be shown from Lemma 2.1.

LEMMA 2.2. *For* $1\leqq k\leqq r$,

$$\sum_{j=1}^{k}\mathrm{ord}_j\,P=-\delta_k(P). \qquad \square$$

In the particular case where $P(s)$ is given as

$$P(s)=C(sF-A)^{-1}B$$

with nonsingular $sF-A$, the above lemma can be formulated as follows.

LEMMA 2.3. *For* $1\leqq k\leqq r$,

$$\sum_{j=1}^{k}\mathrm{ord}_j\,(C(sF-A)^{-1}B)=-\delta_{n+k}\left(\begin{pmatrix}A-sF & B\\ C & O\end{pmatrix};I_0,J_0\right)+\delta(A-sF),$$

where I_0 and J_0 are, respectively, the row and column sets corresponding to the $n\times n$ nonsingular submatrix $A-sF$. □

The following lemma can be shown using Lemma 2.1 and Lemma 2.2 (see, e.g., [5], [24], [32]).

LEMMA 2.4. *Let $P(s)$ and $R(s)$ be proper rational matrices with the same number of rows. There exists a proper rational matrix $X(s)$ such that $P(s)X(s)=R(s)$ if and only if*

$$\operatorname{ord} P(s)=\operatorname{ord}(P(s)\,|\,R(s)). \tag{6}$$ □

Note that (6) implies

$$\operatorname{rank} P(s)=\operatorname{rank}(P(s)\,|\,R(s)).$$

2.2. Mixed matrices. Let $\mathbf{K}$ be a subfield of a field $\mathbf{F}$. A matrix D is called a *mixed matrix* with respect to $\mathbf{F}/\mathbf{K}$ if

$$D=Q+T,$$

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 of its nonzero entries is algebraically independent over $\mathbf{K}$.

(In this paper, the entries of the matrix Q will represent fixed constants, whereas the entries of T will denote independent system parameters. A concrete example of mixed matrix is given in § 6. See Murota [18], [21] for more background about this section.)

The following identity due to Murota and Iri [23] is fundamental. It can be translated nicely into the matroid-theoretic language and enables us to compute the rank of D by an efficient matroid-theoretic algorithm using arithmetic operations in the subfield $\mathbf{K}$ only. Recall that the term-rank of T is defined as the maximum size of a submatrix $T[I,J]$ for which there exists a one-to-one correspondence (permutation) $\pi: I\to J$ such that $T_{i\pi(i)}\neq 0$ for all $i\in I$, i.e.,

$$\text{term-rank } T=\max\{\tau\,|\,\tau=|I|=|J|,\ \exists\pi(\text{one-to-one}): I\to J, \forall i\in I,\ T_{i\pi(i)}\neq 0\}.$$

LEMMA 2.5. *For a mixed matrix $D=Q+T$,*

$$\operatorname{rank} D=\max\{\operatorname{rank} Q[R-I, C-J]+\text{term-rank } T[I,J]\,|\,I\subseteq R, J\subseteq C\},$$

where R and C are the row set and the column set of D. □

Let $\mathbf{K}_0\subseteq \mathbf{F}_0$ be fields, and

$$D(s)=Q(s)+T(s) \tag{7}$$

be a polynomial matrix such that

(i) $Q(s)$ is a polynomial matrix with coefficients from $\mathbf{K}_0$, and

(ii) The set of nonzero coefficients of the entries of $T(s)$ is algebraically independent over $\mathbf{K}_0$.

Then $D(s)$ is a mixed matrix with respect to $\mathbf{F}_0(s)/\mathbf{K}_0(s)$. The following identity, noted by Murota [17], [18], is an extension of Lemma 2.5. Recall that $\delta(\cdot)$ is defined as in § 2.1.

LEMMA 2.6. *Let $D(s)=Q(s)+T(s)$ be a square polynomial mixed matrix as above. Then*

$$\delta(D)=\max\{\delta(Q[R-I, C-J])+\delta(T[I,J])\,|\,|I|=|J|, I\subseteq R, J\subseteq C\},$$

where R and C are the row set and the column set of D.

Proof. The expansion

$$\det(D) = \sum \{\pm \det Q[R-I, C-J] \cdot \det T[I, J] \,|\, |I| = |J|\}$$

shows that

$$\delta(D) \leqq \max \{\delta(Q[R-I, C-J]) + \delta(T[I, J]) \,|\, |I| = |J|\}.$$

Conversely, let $(I, J) = (I^*, J^*)$ yield the maximum on the right-hand side, and consider a term (monomial in the nonzero coefficients in $T[I^*, J^*]$) in $\det T[I^*, J^*]$. The algebraic independence of the nonzero coefficients in T guarantees that this term is not cancelled out but remains in $\det D$. Hence, the above inequality is in fact an equality. □

2.3. Independent assignment. Let $G = (V, E)$ be a directed graph with vertex set V and edge set E. The initial and the terminal vertex of an edge $e \in E$ are denoted by $\partial^+ e$ and $\partial^- e$. $G = (V, E)$ is called a (directed) *bipartite graph* if the vertex set V is partitioned into two disjoint parts as $V = V^+ \cup V^-$ in such a way that all edges are directed from V^+ to V^- (i.e., $\partial^+ e \in V^+$ and $\partial^- e \in V^-$). A *matching* is a subset M of E such that $|M| = |\partial^+ M| = |\partial^- M|$, where $\partial^+ M = \{\partial^+ e \,|\, e \in M\}$, etc. See, e.g., [16] for more about matchings.

A *matroid* is a pair $\mathbf{M} = (U, \mathcal{B})$ of a finite set U and a nonempty collection $\mathcal{B}$ of subsets of U such that:

> If $B_1, B_2 \in \mathcal{B}$ and $b_1 \in B_1 - B_2$, then there exists $b_2 \in B_2 - B_1$ such that $(B_1 \cup \{b_2\}) - \{b_1\} \in \mathcal{B}$.

A subset of U is called a *base* if it belongs to $\mathcal{B}$. All bases of $\mathbf{M}$ have an equal cardinality, which is called the *rank* of $\mathbf{M}$. A subset of U is said to be *independent* if it is contained in a base. See, e.g., [14] and [29] for more about matroids.

Suppose that a bipartite graph $G = (V, E)$ is given and, furthermore, that two matroids, say $\mathbf{M}^+$ and $\mathbf{M}^-$, are defined on V^+ and V^- in terms of the families of bases $\mathcal{B}^+$ and $\mathcal{B}^-$. A matching M is an *independent matching* if $\partial^+ M$ and $\partial^- M$ are independent in $\mathbf{M}^+$ and $\mathbf{M}^-$, respectively. A matching M is an *independent assignment* if $\partial^+ M \in \mathcal{B}^+$ and $\partial^- M \in \mathcal{B}^-$.

Suppose further that an integer weight $\zeta(e)$ is given for each edge $e \in E$, i.e., $\zeta: E \to \mathbf{Z}$. The weight of a matching M is defined by $\zeta(M) = \sum_{e \in M} \zeta(e)$. The independent assignment problem (IAP) is to find an independent assignment M with maximum weight. It is known that IAP is equivalent to the weighted matroid intersection problem. There are a number of very efficient algorithms for finding a maximum independent assignment. See, e.g., [7], [8], [11], [13], [14], and [29] for more about IAP (or weighted matroid intersection problem) and algorithms, and [10], [18] for its applications to engineering problems.

3. Solvability of disturbance decoupling. In this section we summarize some of the standard results on the solvability of DDP. We start by considering a problem that is very much related to DDP. This problem, called the modified disturbance decoupling problem (MDDP), for system (1) consists of finding matrices K and H such that with the feedback $\mathbf{u}(t) = K\mathbf{x}(t) + H\mathbf{d}(t)$ the closed-loop system satisfies

$$C(sI - A - BK)^{-1}(G + BH) = 0. \tag{8}$$

For a descriptor system (4) with nonsingular F, versions of the two decoupling problems DDP and MDDP can be formulated in a straightforward way by observing that descriptor systems (4) with nonsingular F can be transformed into systems of type (1) by a premultiplication with F^{-1}. In that case, DDP for (4) amounts to finding a matrix K such that

$$C(sF-A-BK)^{-1}G=0 \tag{9}$$

and MDDP to finding matrices K and H such that

$$C(sF-A-BK)^{-1}(G+BH)=0. \tag{10}$$

It can be shown with the results of Emre and Hautus [5] (see also Bhattacharyya, Gomes, and Howze [3, Thm. 3], Hautus [9]) that DDP (respectively, MDDP) then is solvable if and only if there exists a strictly proper (respectively, proper) rational matrix $X(s)$ such that

$$C(sF-A)^{-1}BX(s)=C(sF-A)^{-1}G. \tag{11}$$

It then also follows from Lemma 2.4 that MDDP is solvable if and only if

$$\operatorname{ord} C(sF-A)^{-1}B=\operatorname{ord} C(sF-A)^{-1}(B\,|\,G). \tag{12}$$

Noting that

$$\operatorname{rank} C(sF-A)^{-1}B=\operatorname{rank} C(sF-A)^{-1}(B\,|\,G) \tag{13}$$

is equivalent to

$$\operatorname{rank}\begin{pmatrix} A-sF & B\\ C & O\end{pmatrix}=\operatorname{rank}\begin{pmatrix} A-sF & B & G\\ C & O & O\end{pmatrix}, \tag{14}$$

and using Lemma 2.3, we see that (12) holds if and only if

$$\delta_{n+k}\left(\begin{pmatrix} A-sF & B\\ C & O\end{pmatrix}; I_0, J_0\right)=\delta_{n+k}\left(\begin{pmatrix} A-sF & B & G\\ C & O & O\end{pmatrix}; I_0, J_0\right) \tag{15}$$

for $k=1,\cdots,\min(m,p)$, where I_0 and J_0 are, respectively, the rows and columns corresponding to the $n\times n$ submatrix $A-sF$. For later reference we define

$$D(s)=\begin{pmatrix} A-sF & B & G\\ C & O & O\end{pmatrix}. \tag{16}$$

The solvability condition for DDP is obtained from the observation that there exists a strictly proper rational matrix $X(s)$ satisfying (11) if and only if there exists a proper rational matrix $X(s)$ satisfying (11) with G replaced by sG. That is, DDP is solvable if and only if

$$\delta_{n+k}\left(\begin{pmatrix} A-sF & B\\ C & O\end{pmatrix}; I_0, J_0\right)=\delta_{n+k}\left(\begin{pmatrix} A-sF & B & sG\\ C & O & O\end{pmatrix}; I_0, J_0\right) \tag{17}$$

for $k=1,\cdots,\min(m,p)$.

As an extension of (3) we will say that EMMP is solvable for (4) (possibly with singular F) if there exists a strictly proper rational matrix $X(s)$ such that (11) holds.

Summarizing, we can state the following.

LEMMA 3.1. (1) *The* MDDP *for* (4) *with nonsingular F is solvable if and only if* (15) *holds true for* $k=1,\cdots,\min(m,p)$.

(2) *The* DDP *for* (4) *with nonsingular F is solvable if and only if* (17) *holds true for* $k=1,\cdots,\min(m,p)$.

(3) *The* EMMP *for* (4) (*possibly with singular F*) *is solvable if and only if* (17) *holds true for* $k=1,\cdots,\min(m,p)$. □

In this paper we restrict ourselves to nonsingular descriptor systems when we consider DDP or MDDP. DDP for singular systems (with singular F) has been investigated recently by Banaszuk [1], Banaszuk, Kociecki, and Przyluski [2], and Fletcher and Aasaraai [6].

4. Structured system. In this section we give a mathematical formulation for the generic solvability of DDP and the generic structure at infinity for descriptor systems (4) using the structural framework of Murota [17], [18]. First, we briefly summarize two physical observations relevant to analysis of combinatorial structures and describe a mathematical model that represents the combinatorial structure of a system fairly well.

The standard state space representation (1) of a dynamical system has been useful for investigating analytical and algebraic properties of linear systems. Also the investigations of structural properties of linear systems until recently have been mainly based on systems of the type (1). However, as mentioned before in § 1, it has been gradually recognized that systems of type (1) are not powerful enough for the modeling of general physical systems. The reason for this is that due to the incidence structure of a physical system, some of the variables describing the system may be algebraically related to one another. Such relations cannot be incorporated in system descriptions of the form (1), but can be incorporated in those of the form (4). Therefore, systems of the type (4) form a more general framework for the investigation of structural properties of physical systems.

The mathematical model adopted here is based on two different physical observations. One is the distinction between "accurate" and "inaccurate" numbers, and the other is the consistency with respect to physical dimensions.

The first observation, due to Murota and Iri [23], is concerned with how we recognize the structure of a system. When a system is written in the form of (4) in terms of elementary variables, it is often justified to assume that the nonzero entries of the matrices F, A, etc., are classified into two groups. One group of generic parameters and the other group of fixed constants. In other words, we can distinguish the following two kinds of numbers, together characterizing a physical system:

1) Inaccurate numbers. Numbers representing independent physical parameters such as resistances in electrical networks, which, being contaminated with noise and other errors, take values independent of one another, and therefore can be modeled as algebraically independent numbers; and
2) Accurate numbers. Numbers accounting for various sorts of conservation laws such as Kirchhoff's laws, which, stemming from topological incidence relations, are precise in value (often ±1), and therefore cause no serious numerical difficulty in arithmetic operations on them.

We may also refer to the numbers of the first kind as "system parameters" and to those of the second kind as "fixed constants."

This observation can be translated into a mathematical assumption using the notion of mixed matrices. That is, we assume that the matrices F, A, etc., in (4) are mixed matrices with respect to $\mathbf{F}_0/\mathbf{Q}$ (where $\mathbf{F}_0$ is a sufficiently large field and $\mathbf{Q}$ the field of rational numbers) expressed as

$$F = Q_F + T_F, \qquad A = Q_A + T_A, \quad \text{etc.},$$

where Q_F, Q_A, etc., are matrices over $\mathbf{Q}$, and, furthermore, we assume that

(A1) The collection of nonzero entries of T_F, T_A, etc., are algebraically independent over $\mathbf{Q}$.

Then $D(s)$ of (16) is a mixed matrix with respect to $\mathbf{F}_0(s)/\mathbf{Q}(s)$, accordingly expressed as

$$D(s)=Q_D(s)+T_D(s) \tag{18}$$

with

$$Q_D(s)=\begin{pmatrix} Q_A-sQ_F & Q_B & Q_G \\ Q_C & O & O \end{pmatrix}, \qquad T_D(s)=\begin{pmatrix} T_A-sT_F & T_B & T_G \\ T_C & O & O \end{pmatrix}.$$

The second observation, due to Murota [17], is concerned with the "accurate numbers," i.e., with $Q_D(s)$ in (18). The "accurate numbers" usually represent topological and/or geometrical incidence coefficients, which have no physical dimensions, so that it is natural to expect that the entries of Q_F, Q_A, 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.

Since the system (4) is to represent a physical system, relevant physical dimensions are associated with both the variables and the equations, or alternatively, with the columns and the rows of $D(s)$. Choosing time as one of the fundamental dimensions, we denote by $-c_j$ and $-r_i$ the exponent to the dimension of time associated, respectively, with the jth column and the ith row. The principle of dimensional homogeneity then requires that the (i,j) entry of $D(s)$ should have the dimension of time with exponent c_j-r_i.

Combining this fact with the observations on the nondimensionality of Q_F, Q_A, etc., and on the dimension of s, we obtain

$$Q_D(s)=\operatorname{diag}(s^{r_1},\cdots,s^{r_{n+p}})\cdot Q_D(1)\cdot\operatorname{diag}(s^{-c_1},\cdots,s^{-c_{n+m+l}}). \tag{19}$$

It can be shown by purely linear-algebraic arguments without reference to physical dimensions (cf. [17], [18]) that a polynomial matrix $Q_D(s)$ can be represented as (19) with some (nonunique) r_i and c_j if and only if

(A2) Every nonzero subdeterminant of $Q_D(s)$ is a monomial in s over $\mathbf{Q}$.

Hence, the principle of dimensional homogeneity comes down to assuming (A2). It should be emphasized that this algebraic assumption (A2) expresses the physical-dimensional consistency among the "accurate numbers" or "fixed constants."

The mathematical model used here for structural analysis consists of (4) or (16), which satisfy (A1) and (A2). Adopting this model, we will give combinatorial characterizations of the following:

1) The properness of the transfer matrix of a descriptor system (possibly with singular F), i.e., $\delta_1(C(sF-A)^{-1}B)$;
2) The structure at infinity of the transfer matrix of a descriptor system (possibly with singular F), i.e., $\operatorname{ord}(C(sF-A)^{-1}B)$;
3) The solvability of DDP and MDDP for a descriptor system with nonsingular F;
4) The solvability of EMMP for a descriptor system (possibly with singular F).

It should be obvious that the special case of the present model with $Q_D(s)=0$ reduces to the conventional framework of a structured descriptor system in which all the nonzero entries are assumed to be independent parameters; note that (A2) is satisfied trivially when $Q_D(s)=0$.

From Lemma 2.3 and Lemma 3.1 it follows that the characterizations above can be obtained by solving the problem below. Note that a combinatorial characterization (cf. [18, § 19]) derived from Lemma 2.5 for the rank of a mixed matrix is already available.

PROBLEM. Find a combinatorial characterization to $\delta_k(D; I_0, J_0)$ for a mixed matrix $D(s)$ of the form (7) (not necessarily of the form (16)) such that $Q(s)$ satisfies (A2), where I_0 and J_0 are specified row and column subsets, and $k \geqq \max(|I_0|, |J_0|)$.

Remark 4.1. In this paper we consider models for which among others, assumption (A1) is satisfied. Denoting by $\mathcal{T}$ the collection of the nonzero entries of T_F, T_A, etc., it is clear that assumption (A1) on algebraic independence of $\mathcal{T}$ is introduced to investigate "generic properties" of the family of descriptor systems parametrized by $\mathcal{T}$.

The four properties listed above are in fact generic properties with respect to this parametrization $\mathcal{T}$. For example, the solvability of DDP is characterized in terms of the degrees of minors in s, as stated in Lemma 3.1(2), and the degree of each minor remains constant outside an algebraic variety in the space of $\mathcal{T}$. This shows that the solvability of DDP is a generic property.

Hence, our combinatorial characterizations to be obtained are valid for almost all parameter values of $\mathcal{T}$, and can fail only for "combinations" of the values of $\mathcal{T}$ that are algebraically dependent over $\mathbf{Q}$. To reiterate, our results yield "generic" conditions for the system-theoretic properties, where the "genericity" is defined with respect to $\mathcal{T}$.

There is another point to be noted about the significance of the generic solvability of DDP and MDDP. Even though Lemma 3.1(2) may hold for a particular system, this does not mean that the generic value of the left-hand side of (17) is equal to the generic value of the right-hand side of (17), for $k = 1, \cdots, \min(m, p)$. This is because the equalities may be satisfied, and hence DDP may be solvable, for a special combination of the parameter values of $\mathcal{T}$ that are algebraically dependent over $\mathbf{Q}$, while generically at least one of the equalities is violated. Thus, the usual solvability of DDP does not imply the generic solvability of DDP. The same holds of course for the solvability of MDDP. Note that this stands in sharp contrast to the structural controllability. Namely, if a system is controllable in the usual sense, then the system must be structurally controllable.

5. Combinatorial characterization.

In this section we give a solution of the following problem.

PROBLEM. Find a combinatorial characterization to $\delta_k(D; I_0, J_0)$ (cf. (5) for notation) for a mixed matrix $D(s) = Q(s) + T(s)$ as in (7) such that $Q(s)$ satisfies (A2), i.e.,

$$Q(s) = \operatorname{diag}(s^{r_i} \mid i \in R) \cdot Q(1) \cdot \operatorname{diag}(s^{-c_j} \mid j \in C), \tag{20}$$

for some $r_i \in \mathbf{Z}$ and $c_j \in \mathbf{Z}$, where R and C denote the row and the column set of D, and where $I_0 \subseteq R$, $J_0 \subseteq C$ and $k \geqq \max(|I_0|, |J_0|)$ are specified. □

Before considering the general case we explain our approach for the special case that $I_0 = J_0 = \emptyset$. That is, we will first consider a combinatorial characterization for $\delta_k(D) = \delta_k(D; \emptyset, \emptyset)$. After that we indicate how the obtained result can be extended to the general case.

First note that Lemma 2.6 for $\delta(D)$ can be extended for $\delta_k(D)$ as follows.

LEMMA 5.1. *Let $D(s) = Q(s) + T(s)$ be a polynomial mixed matrix as above. Then*

$$\begin{aligned}\delta_k(D) = \max\{\delta(Q[I_1, J_1]) + \delta(T[I_2, J_2]) \mid\ & |I_i| = |J_i|, I_i \subseteq R, J_i \subseteq C\ (i = 1, 2); \\ & |I_1| + |I_2| = k, I_1 \cap I_2 = J_1 \cap J_2 = \emptyset\},\end{aligned}$$

where R and C are the row set and the column set of D.

Proof. By definition,

$$\delta_k(D) = \max\{\delta(D[I,J]) \mid |I| = |J| = k\}.$$

The claim follows immediately from Lemma 2.6 applied to $\delta(D[I,J])$. □

We will reformulate Lemma 5.1 above as an independent assignment problem (IAP). Note for $k \leqq \operatorname{rank} D$ there exists a tuple $(I_1, J_1; I_2, J_2)$ as above such that $Q[I_1, J_1]$ and $T[I_2, J_2]$ are nonsingular, where, for example, $Q[\emptyset, \emptyset]$ is assumed to be nonsingular. Suppose the maximum in Lemma 5.1 is attained by $(I_1, J_1; I_2, J_2)$.

We associate with D a bipartite graph $G = G(D) = (V, E)$ having vertex bipartition $V = V^+ \cup V^-$ with

$$V^+ = R_T \cup R_Q \cup C_Q, \qquad V^- = R \cup C,$$

where R_T and R_Q are disjoint copies of R, and C_Q is a disjoint copy of C. By $\varphi_Q : R \cup C \to R_Q \cup C_Q$ and $\varphi_T : R \to R_T$ will denote the natural correspondences between the copies. The edge set E is defined by

$$\begin{aligned} E = &\{(\varphi_Q(i), i) \mid i \in R\} \cup \{(\varphi_Q(j), j) \mid j \in C\} \\ &\cup \{(\varphi_T(i), i) \mid i \in R\} \cup \{(\varphi_T(i), j) \mid T_{ij} \neq 0, i \in R, j \in C\}. \end{aligned}$$

Since $T[I_2, J_2]$ is nonsingular, we may associate with $(I_1, J_1; I_2, J_2)$ a matching M in G defined by

$$\begin{aligned} M = &\{(\varphi_Q(i), i) \mid i \in R - I_1\} \cup \{(\varphi_Q(j), j) \mid j \in J_1\} \\ &\cup \{(\varphi_T(i), i) \mid i \in I_1\} \cup \{(\varphi_T(i), j) \mid i \in I_2, j \in J_2, j = \pi(i)\}, \end{aligned} \tag{21}$$

where π is a permutation $\pi : I_2 \to J_2$ such that $T_{i\pi(i)} \neq 0$ (for all $i \in I_2$). Then

$$\begin{aligned} \partial^+ M &= (R_Q - \varphi_Q(I_1)) \cup \varphi_Q(J_1) \cup \varphi_T(I_1 \cup I_2), \\ \partial^- M &= R \cup J_1 \cup J_2. \end{aligned}$$

We now show that M is an independent assignment in G with respect to matroids appropriately defined on V^+ and V^-. First we define

$$\mathscr{L} = \{(I_1, J_1) \mid Q[I_1, J_1] \text{ is nonsingular}, I_1 \subseteq R, J_1 \subseteq C\},$$
$$\mathscr{L}_Q = \{(\varphi_Q(I_1), \varphi_Q(J_1)) \mid (I_1, J_1) \in \mathscr{L}\}$$

and next we consider two families of subsets of V^+ and V^-, respectively, defined as

$$\mathscr{B}^+ = \mathscr{B}_k^+ = \{U^+ \subseteq V^+ \mid (R_Q - U^+, C_Q \cap U^+) \in \mathscr{L}_Q, |U^+| = |R| + k\},$$
$$\mathscr{B}^- = \mathscr{B}_k^- = \{U^- \subseteq V^- \mid U^- \supseteq R, |U^-| = |R| + k\}.$$

These two families, $\mathscr{B}^+$ and $\mathscr{B}^-$, define matroids, as claimed below.

LEMMA 5.2. (1) *$\mathscr{B}_k^+$ forms a base of a matroid, say $\mathbf{M}^+ = \mathbf{M}_k^+$, of rank $|R| + k$; $\mathbf{M}^+$ is the direct sum of a linear matroid of rank $|R|$ representable over $\mathbf{K}_0$ and a uniform matroid of rank k.*

(2) *$\mathscr{B}_k^-$ forms a base family of a matroid, say $\mathbf{M}^- = \mathbf{M}_k^-$, of rank $|R| + k$; $\mathbf{M}^-$ is the direct sum of a uniform matroid of rank k and a free matroid of rank $|R|$.*

Proof. (1) Put $I_1 = \varphi_Q^{-1}(R_Q - U^+)$ and $J_1 = \varphi_Q^{-1}(C_Q \cap U^+)$ and note that $Q[I_1, J_1]$ is nonsingular if and only if the column vectors of the compound matrix $\tilde{Q}(s) = (I \mid Q(s))$ corresponding to $(R - I_1) \cup J_1$ form a basis, where the column set of $\tilde{Q}(s)$ is identified with $R \cup C$. Observe further that the assumption (A2) implies that columns of $\tilde{Q}(s)$ are independent if and only if the corresponding columns of $\tilde{Q}(1)$, which is a matrix over $\mathbf{K}_0$, are also independent. This shows that the condition $(R_Q - U^+, C_Q \cap U^+) \in \mathscr{L}_Q$ in the definition of $\mathscr{B}^+$ defines a linear matroid on $R_Q \cup C_Q$ that is representable over $\mathbf{K}_0$. The second condition, being equivalent to $|R_T \cap U^+| = k$, defines a uniform matroid on R_T of rank k. Hence, $\mathbf{M}^+$ is the direct sum of the two matroids as claimed.

(2) Similarly, $\mathbf{M}^-$ is the direct sum of a free matroid on R and a uniform matroid of rank k on C. □

Since $\partial^+ M \in \mathcal{B}^+$ and $\partial^- M \in \mathcal{B}^-$ by the construction, we see that the matching M associated with $(I_1, J_1; I_2, J_2)$ is in fact an independent assignment in G with respect to $\mathbf{M}^+$ and $\mathbf{M}^-$.

Conversely, an independent assignment M determines a tuple $(I_1, J_1; I_2, J_2)$ such that $I_1 \cap I_2 = \emptyset$, $J_1 \cap J_2 = \emptyset$, and both $Q[I_1, J_1]$ and $T[I_2, J_2]$ are nonsingular. In fact, these subsets are determined by the following relations:

$$I_1 = \varphi_Q^{-1}(R_Q - \partial^+ M), \qquad J_1 = \varphi_Q^{-1}(C_Q \cap \partial^+ M),$$
$$I_2 = \varphi_T^{-1}(R_T \cap \partial^+ M) - I_1, \qquad J_2 = (C \cap \partial^- M) - J_1.$$

Next we will introduce a weight function $\zeta : E \to \mathbf{Z}$ with reference to the degrees of the entries of $Q(s)$ and $T(s)$. Using the numbers r_i and c_j in (20), we define, for $e \in E$,

$$\zeta(e) = \begin{cases} -r_i & \text{if } e = (\varphi_Q(i), i), \quad i \in R, \\ -c_j & \text{if } e = (\varphi_Q(j), j), \quad j \in C, \\ \deg_s T_{ij} & \text{if } e = (\varphi_T(i), j), \quad i \in R, \quad j \in C, \\ 0 & \text{if } e = (\varphi_T(i), i), \quad i \in R. \end{cases} \tag{22}$$

Then from (20)-(22), we see that

$$\begin{aligned} \zeta(M) &= -\sum_{i \in R - I_1} r_i - \sum_{j \in J_1} c_j + \sum_{i \in I_2} \deg_s T_{i\pi(i)} \\ &= \delta(Q[I_1, J_1]) - r_0 + \sum_{i \in I_2} \deg_s T_{i\pi(i)}, \end{aligned}$$

where

$$r_0 = \sum_{i \in R} r_i. \tag{23}$$

In addition we have

$$\sum_{i \in I_2} \deg_s T_{i\pi(i)} = \delta(T[I_2, J_2])$$

for an appropriate choice of π. Therefore, we have the relation

$$\max_M \zeta(M) = \max_{(I_1, J_1; I_2, J_2)} \{\delta(Q[I_1, J_1]) + \delta(T[I_2, J_2])\} - r_0,$$

where the maximum on the left-hand side is taken over all independent assignments M, and that on the right-hand side is over all tuples $(I_1, J_1; I_2, J_2)$ having the properties stated before. Combining the above identity with Lemma 5.1, we obtain the combinatorial characterization for $\delta_k(D)$. In the case where $k = |R| = |C|$, this result reduces (essentially) to the result for $\delta(D)$ obtained by Murota [17], [18].

THEOREM 5.1. *For $D(s)$ of* (7) *having property* (20) *and for $k \geqq 0$,*

$$\delta_k(D) = \max_M \zeta(M) + r_0,$$

where the maximum on the right-hand side is taken over all independent assignments M in the bipartite graph $G(D)$ with matroids defined by $\mathcal{B}_k^+$ and $\mathcal{B}_k^-$, and r_0 is defined by (23). □

The ideas expounded above for the case that $I_0 = J_0 = \emptyset$ can be extended to the general case. To this end, we consider an IAP on the same graph $G(D)$ but with slightly different matroids $\mathbf{M}^+ = \mathbf{M}_k^+(I_0)$ and $\mathbf{M}^- = \mathbf{M}_k^-(J_0)$ defined by

$$\begin{aligned} \mathcal{B}^+ = \mathcal{B}_k^+(I_0) = \{U^+ \subseteq V^+ \mid (R_Q - U^+, C_Q \cap U^+) \in \mathcal{L}_Q, \\ R_T \cap U^+ \supseteq \varphi_T(I_0), |U^+| = |R| + k\}, \end{aligned}$$
$$\mathcal{B}^- = \mathcal{B}_k^-(J_0) = \{U^- \subseteq V^- \mid U^- \supseteq R \cup J_0, |U^-| = |R| + k\}.$$

In the spirit of Lemma 5.2 we can prove that $\mathbf{M}^+$ is the direct sum of a linear matroid of rank $|R|$ representable over $\mathbf{K}_0$, a uniform matroid of rank $k-|I_0|$ and a free matroid of rank $|I_0|$, whereas $\mathbf{M}^-$ is the direct sum of a uniform matroid of rank $k-|J_0|$ and a free matroid of rank $|R|+|J_0|$. The weight function ζ remains unchanged. The main result of this paper is now stated.

THEOREM 5.2. *For* $D(s)$ *of* (7) *having property* (20) *and for* $I_0 \subseteq R$, $J_0 \subseteq C$ *and* $k \geqq \max(|I_0|, |J_0|)$ *as specified before,*

$$\delta_k(D; I_0, J_0) = \max_M \zeta(M) + r_0,$$

where the maximum on the right-hand side is taken over all independent assignments M *in the bipartite graph* $G(D)$ *with matroids defined by* $\mathcal{B}_k^+(I_0)$ *and* $\mathcal{B}_k^-(J_0)$, *and* r_0 *is defined by* (23). □

The established characterization provides us with an efficient and practical way of computing $\delta_k(D; I_0, J_0)$ by means of well-established algorithms for the independent assignment problem (see [7], [8], [11], [13], and [14] for algorithms). It then follows from the arguments in § 4 that we have efficient algorithms for checking the properness of a transfer function matrix, for computing the structure at infinity of a transfer function matrix, and for testing for the solvability of EMMP, DDP, and MDDP. Remember, however, that the algorithms yield "generic" results for descriptor systems, where the "genericity" is defined with respect to the parametrization by the nonzero entries of the T-part (cf. Remark 4.1).

In an application of the above result to the structure at infinity, we have to compute $\delta_k(D; I_0, J_0)$ for $k = n+1, \cdots, n+r$ with D, I_0, and J_0 fixed (see Lemma 2.3). This means that we have to solve r problems in the same bipartite graph, but with different matroids $\mathbf{M}_k^+(I_0)$ and $\mathbf{M}_k^-(J_0)$ for $k = n+1, \cdots, n+r$. However, these matroids are closely related to one another in the sense that an independent matching for $k = k_1$ is also an independent matching for $k = k_1 + 1$. In the standard augmenting algorithm for IAP, we can therefore use the maximum weight independent assignment M for $k = k_1$ as the starting independent matching for the problem with $k = k_1 + 1$; M is to be increased by one in size.

In another application of the above result to the solvability of EMMP, DDP, or MDDP we have to compare $\delta_k(D_1; I_0, J_0)$ with $\delta_k((D_1|D_2); I_0, J_0)$ for each k (see [15] and [17]). Here we may regard the graph $G(D_1)$ as a subgraph of $G((D_1|D_2))$ and an independent assignment in $G(D_1)$ is also an independent assignment in $G((D_1|D_2))$. This means that, once a maximum weight independent assignment is found in $G(D_1)$, the test for the equality $\delta_k(D_1; I_0, J_0) = \delta_k((D_1|D_2); I_0, J_0)$ can be done simply by detecting a negative-length cycle in the auxiliary graph used in standard implementations of the algorithms for IAP.

A special case of Theorem 5.1 is used in van der Woude [33] in the computation of a splitting of the external variable of a linear structured system in AR-form into an input and an output in such a way that the output depends on the input in a nonanticipative way (cf. Willems [30]).

A variant of Theorem 5.2 is used by Murota [22] in designing an efficient algorithm for computing the Smith normal form of a polynomial matrix $D(s)$ of (7) having property (20).

6. Example. This section presents an example that illustrates the method proposed in § 5 in comparison with the graph-theoretic method of van der Woude [32] (and also Commault, Dion, and Perez [4] and Suda, Wan, and Ueno [27]). We will compute the structure at infinity of the transfer matrix $P(s) = C(sF-A)^{-1}B$ of a descriptor system (4) ($n=6$, $m=2$, $p=2$) defined by

$$A-sF=\begin{pmatrix} -s & 0 & 1 & 0 & 0 & 0 \\ 0 & -s & 0 & 1 & 0 & 0 \\ -k_1 & 0 & -sm_1 & 0 & -1 & 0 \\ 0 & -k_2 & 0 & -sm_2 & 1 & 0 \\ 0 & 0 & 0 & 0 & -1 & f \\ -s & s & 0 & 0 & 0 & 1 \end{pmatrix},$$

$$B=\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \qquad C=\begin{pmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \end{pmatrix}.$$

We regard $\{m_1, m_2, k_1, k_2, f\}$ as independent free parameters. (This example represents a small mechanical system consisting of two masses m_1 and m_2, two springs k_1 and k_2, and one damper f (see [18, § 18]).) We have the mixed matrix

$$D(s)=\begin{pmatrix} -s & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -s & 0 & 1 & 0 & 0 & 0 & 0 \\ -k_1 & 0 & -sm_1 & 0 & -1 & 0 & 1 & 0 \\ 0 & -k_2 & 0 & -sm_2 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & f & 0 & 0 \\ -s & s & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \end{pmatrix},$$

which can be decomposed as (18) with

$$Q_D(s)=\begin{pmatrix} -s & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -s & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ -s & s & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

and

$$T_D(s)=\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -k_1 & 0 & -sm_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -k_2 & 0 & -sm_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & f & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Note that $Q_D(s)$ admits an expression of the form (20) with

$$r_1=r_2=1,\quad r_3=r_4=r_5=2,\quad r_6=1,\quad r_7=0,\quad r_8=1,$$
$$c_1=c_2=0,\quad c_3=c_4=1,\quad c_5=2,\quad c_6=1,\quad c_7=c_8=2.$$

(These numbers are determined from the physical dimensions associated with variables and equations, although they are not unique from the mathematical point of view (see [18, § 18]).)

We denote the rows of D by w_i $(i=1,\cdots,6)$ and y_i $(i=1,2)$, and the columns by x_j $(j=1,\cdots,6)$ and u_j $(j=1,2)$. That is,

$$R=\{w_i\,|\,i=1,\cdots,6\}\cup\{y_1,y_2\},$$
$$C=\{x_j\,|\,j=1,\cdots,6\}\cup\{u_1,u_2\}.$$

The subsets I_0 and J_0 specified in Lemma 2.3 are given by

$$I_0=\{w_i\,|\,i=1,\cdots,6\},\qquad J_0=\{x_j\,|\,j=1,\cdots,6\}.$$

The associated graph $G(D)=(V,E)$ has 40 vertices ($|V^+|=24$, $|V^-|=16$) and 29 edges with weight ζ of (22) (see Fig. 1). Among those edges, 24 edges represent the correspondence between copies, and the remaining 5 edges denote the independent parameters as listed below, where we use the short-hand notation $\varphi_T(w_i)=w_i^T$:

parameters:	m_1	m_2	k_1	k_2	f
edges:	(w_3^T,x_3)	(w_4^T,x_4)	(w_3^T,x_1)	(w_4^T,x_2)	(w_5^T,x_6).
weights:	1	1	0	0	0

By Lemma 2.5 we see that $A-sF$ is in fact nonsingular with rank equal to 6, and, furthermore, by Lemma 2.6 that $\delta(A-sF)=4$. On the other hand, Lemma 2.5 reveals that $D(s)$ is singular with rank $D(s)=7$ although term-rank $D(s)=8$. Hence,

$$\text{rank } P(s)=\text{rank } D(s)-\text{rank }(A-sF)=1.$$

By solving the IAP with matroids $\mathbf{M}_7^+(I_0)$ and $\mathbf{M}_7^-(J_0)$, we find an independent assignment M with maximum weight $\zeta(M)=-7$; for example, such M is given by (21) with $I_1=\{w_1,w_2,w_3,w_5,w_6,y_2\}$, $J_1=\{x_1,x_2,x_3,x_5,x_6,u_1\}$, $I_2=\{w_4\}$, $J_2=\{x_4\}$. Since $r_0=10$, we obtain

$$\delta_7(D(s);\,I_0,J_0)=\zeta(M)+r_0=3.$$

It then follows from Lemma 2.3 that

$$\text{ord}_1\,P(s)=-3+4=1.$$

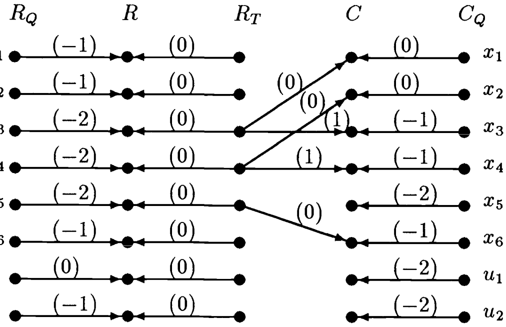


FIG. 1. *Graph $G(D)$ of example* ((·): *weight*).

For comparison let us apply the graph-theoretic method of van der Woude [32] to this example. We first transform this system into the standard state space form (1) to obtain a fourth-order system ($n=4$) with

$$\hat{A}=\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -k_1/m_1 & 0 & -f/m_1 & f/m_1 \\ 0 & -k_2/m_2 & f/m_2 & -f/m_2 \end{pmatrix},$$

$$\hat{B}=\begin{pmatrix} 0 & 0 \\ 0 & 0 \\ 1/m_1 & 0 \\ 0 & 1/m_2 \end{pmatrix}, \qquad \hat{C}=\begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

We denote the transfer matrix of this system by

$$\hat{P}(s)=\hat{C}(sI-\hat{A})^{-1}\hat{B}.$$

Note that $P(s)=\hat{P}(s)$.

The associated digraph $\hat{G}$ is depicted in Fig. 2. As is easily seen, $\hat{G}$ admits a Menger-type vertex-disjoint complete linking (of size 2) from $\{u_1, u_2\}$ to $\{y_1, y_2\}$. This means (see, for instance, [18, Thm. 7.2] and [32, Thm. 6.2]) that if the nonzero entries of $\hat{A}$, $\hat{B}$ and $\hat{C}$ were algebraically independent, then

$$\operatorname{rank} \hat{P}(s)=2.$$

Furthermore, the graph-theoretic criterion of van der Woude [32] tells us that

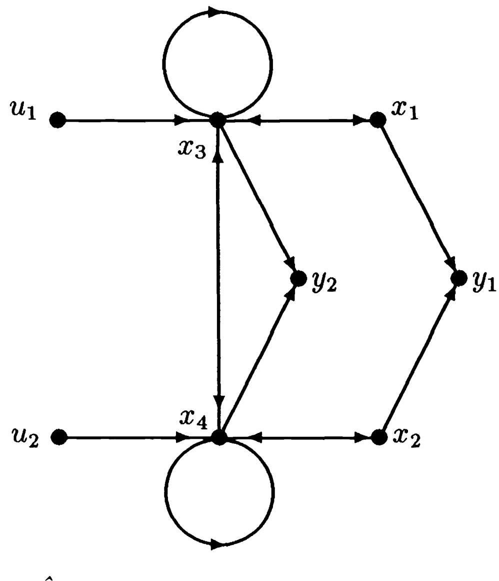


FIG. 2. *Graph $\hat{G}$ of example (↔ stands for a pair of edges in opposite directions).*

$\sum_{j=1}^{k} \operatorname{ord}_j \hat{P}(s)$ is equal to the minimum number of vertices x_j contained in a vertex-disjoint linking of size k from $\{u_1, u_2\}$ to $\{y_1, y_2\}$, and hence that

$$\operatorname{ord}_1 \hat{P}(s) = 1, \qquad \operatorname{ord}_2 \hat{P}(s) = 2,$$

provided again that the nonzero entries of $\hat{A}$, $\hat{B}$, and $\hat{C}$ were algebraically independent. However, in the present example the conclusion that rank $\hat{P}(s) = 2$, and $\operatorname{ord}_1 \hat{P}(s) = 1$, $\operatorname{ord}_2 \hat{P}(s) = 2$, is incorrect because the entries of the matrices are not algebraically independent.

From the previous it is clear that the correct conclusion is that rank $\hat{P}(s) = 1$ and $\operatorname{ord}_1 \hat{P}(s) = 1$. Therefore, in the case of algebraically related entries the structure at infinity of a physical system can be determined better by the matroid-theoretic method developed in this paper than by the graph-theoretic method of [4], [27], or [32].

Acknowledgments. The comments of the anonymous referees were helpful in revision.

REFERENCES

[1] A. Banaszuk, *The disturbance decoupling with stabilization for implicit linear discrete-time systems*, in Proc. IFAC Workshop on System Structure and Control: State-Space and Polynomial Methods, Prague, September 1989, pp. 25-27.

[2] A. Banaszuk, M. Kociecki, and K. M. Przyluski, *The disturbance decoupling problem for implicit linear discrete-time systems*, SIAM J. Control Optim., 28 (1990), pp. 1270-1293.

[3] S. P. Bhattacharyya, A. C. D. N. Gomes, and J. W. Howze, *The structure of robust disturbance rejection control*, IEEE Trans. Automat. Control, 28 (1983), pp. 874-881.

[4] C. Commault, J.-M. Dion, and A. Perez, *Rejet de perturbation dans les systemes structures*, Laboratoire d'Automatique de Grenoble (URA CNRS), 1989. (English version: *Disturbance rejection for structured systems.*)

[5] E. Emre and M. L. J. Hautus, *A polynomial characterization of (A, B)-invariant and reachability subspaces*, SIAM J. Control Optim., 18 (1980), pp. 420-436.

[6] L. R. Fletcher and A. Aasaraai, *On disturbance decoupling in descriptor systems*, SIAM J. Control Optim., 27 (1989), pp. 1319-1332.

[7] H. N. Gabow and Y. Xu, *Efficient theoretic and practical algorithms for linear matroid intersection problems*, CU-CS-424-89, Department of Computer Science, University of Colorado, Boulder, CO, 1989.

[8] ———, *Efficient algorithms for independent assignment on graphic and linear matroids*, in Proc. IEEE 30th Annual Symposium on Foundations of Computer Science, IEEE Computer Society, Washington, DC, November 1989, pp. 106-111.

[9] M. L. J. Hautus, *(A, B)-invariant and stabilizability subspaces, a frequency domain description*, Automatica, 16 (1980), pp. 703-707.

[10] M. Iri, *Applications of matroid theory*, in Mathematical Programming—State of the Art, A. Bachem, M. Grötschel, and B. Korte, eds., Springer-Verlag, Berlin, 1983, pp. 158-201.

[11] M. Iri and N. Tomizawa, *An algorithm for finding optimal "independent assignment,"* J. Oper. Res. Soc. Japan, 19 (1975), pp. 32-57.

[12] N. Kobayashi and T. Nakamizo, *A disturbance rejection problem in structural aspects*, Trans. Soc. Instr. Control Engineers, 23 (1987), pp. 928-934. (In Japanese.)

[13] E. L. Lawler, *Matroid intersection algorithms*, Math. Programming, 9 (1975), pp. 31-56.

[14] ———, *Combinatorial Optimization: Networks and Matroids*, Holt, Reinhart and Winston, New York, 1976.

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

[16] L. Lovász and M. Plummer, *Matching Theory*, North-Holland, Amsterdam, 1986.

[17] 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.

[18] ———, *Systems Analysis by Graphs and Matroids—Structural Solvability and Controllability*, Algorithms and Combinatorics, Vol. 3, Springer-Verlag, Berlin, 1987.

[19] ———, *Refined study on structural controllability of descriptor systems by means of matroids*, SIAM J. Control Optim., 25 (1987), pp. 967-989.

[20] ———, *A matroid-theoretic approach to structurally fixed modes of control systems*, SIAM J. Control Optim., 27 (1989), pp. 1381-1402.

[21] ———, *Some recent results in combinatorial approaches to dynamical systems*, Linear Algebra Appl., 122/123/124 (1989), pp. 725-759.

[22] ———, *On the Smith normal form of structured polynomial matrices*, SIAM J. Matrix Anal. Appl., 12 (1991), to appear.

[23] K. MUROTA AND M. IRI, *Structural solvability of a system 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.

[24] M. NEWMAN, *Integral Matrices*, Academic Press, London, 1972.

[25] K. REINSCHKE, *Struktuelle Pol-Nullstellen-Analyse von Systemen in Zustandsraumdarstellung*, Messen-Steuern-Regeln, 25 (1982), pp. 542-550.

[26] W. SÖTE, *Eine graphische Methode zur Ermittlung der Nullstellen in Mehrgrößensystemen*, Regelungstechnik, 28 (1980), pp. 346-348.

[27] N. SUDA, B. WAN, AND I. UENO, *The orders of infinite zeros of structured systems*, Trans. Soc. Instr. Control Engineers, 25 (1989), pp. 1062-1068. (In Japanese.)

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

[29] D. J. A. WELSH, *Matroid Theory*, Academic Press, New York, 1976.

[30] J. C. WILLEMS, *From time series to linear system—part* I. *Finite dimensional linear time invariant systems*, Automatica, 22 (1986), pp. 561-580.

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

[32] J. W. VAN DER WOUDE, *On the structure at infinity of a structured system*, Linear Algebra Appl., to appear.

[33] ———, *Causality relations in linear structured systems*, Communication, Control and Signal Processing, Proc. 1990 Bilkent Intern. Conference on New Trends in Communication, Control, and Signal Processing, E. Arikan, ed., Elsevier, Amsterdam, 1990, pp. 722-778.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 895-908, July 1991

NECESSARY CONDITIONS FOR OPTIMAL CONTROL OF DISTRIBUTED PARAMETER SYSTEMS*

XUNJING LI† AND JIONGMIN YONG†

Abstract. The Ekeland variational principle is used to prove a maximum principle for the optimal controls of a general semilinear evolutionary control system in a Banach space with strictly convex dual. The boundary constraint is of general type which includes the optimal periodic control problem as a special case.

Key words. maximum principle, Ekeland's variational principle, optimal control, distributed parameter system

AMS(MOS) subject classifications. 49B27, 93C25

1. Introduction. Necessary conditions for optimal control of lumped parameter systems were derived by Pontryagin et al. [29] (see also [6]-[9], [28], [30]). Butkovskii [7] was the first to discuss the optimal control problems for distributed parameter systems. The maximum principle as a set of necessary conditions for optimal control of distributed parameter systems has been studied by many authors. Since it is well known that the maximum principle may be false for distributed parameter systems (see [4]), there are many papers that give some conditions ensuring that the maximum principle remains true. We refer the readers to [1], [4], [5], [11], and [26] for some results of this aspect. We note that the above-mentioned references discuss the cases for distributed parameter systems or functional differential systems with no end constraints and/or with the control domain being convex. Thus, they do not include the Pontryagin original result as a special case. It was Li and Yao [25] who proved a maximum principle, for distributed parameter systems in some Banach spaces with finite-codimensional target set, which covered the Pontryagin result. Fattorini [18], [19] proved a maximum principle for distributed parameter systems in some Hilbert spaces under the condition that the reachable set of the variational system is of finite codimension. Xu [31] proved a similar result in a uniformly convex Banach space.

Problems with general end constraints (meaning that the initial and the terminal states of the system belong to some subset of the product space of the state space; see § 2 for details) for lumped parameter systems were formulated by Berkovitz [6] and he pointed out that, by increasing the dimension of the state space, such problems can be reduced to the one with separated end constraints, studied by Pontryagin et al.

Colonius [10], Li [23], and Li and Chow [24] considered the optimal periodic control problems for functional differential systems. They proved the maximum principle under some technical assumptions. Li and Chow [24] and Jin and Li [22] studied the problems with general end constraints for functional differential systems and derived the necessary conditions for the optimal controls.

In this paper, we use Ekeland's variational principle to prove a maximum principle for the optimal controls of a general semilinear evolutionary control system in some Banach space with general end constraints which contains the periodic optimal control problem as a special case. We should note that unlike the finite-dimensional case, we

* Received by the editors June 5, 1989; accepted for publication (in revised form) August 10, 1990. This work was partially supported by Chinese Natural Science Foundation grants 1880414 and 0188416 and by Chinese State Education Commission Science Foundation grant 8824601.

† Department of Mathematics, Fudan University, Shanghai 200433, China.

could not simply use the argument of [6] to reduce such a problem to the one with separated end constraints.

2. Optimal control problem, maximum principle. In this section, we give the basic framework of our optimal control problem. We will also state the main result of this paper—the maximum principle, the proof of which will be given in § 4.

Let X and Y be Banach spaces with the embedding $X \hookrightarrow Y$ being dense and continuous. We denote the dual spaces of X and Y by X^* and Y^*, respectively. We assume the following:

(H0) X^* is strictly convex.

By [2] and [3] (see also [15]), we know that if X is reflexive, then, by changing the norm to an equivalent one, we may assume that X^* is strictly convex. Also, if X is separable, we may also do the above (see the proof of Theorem 4 of [12], see also [13], [14]). Thus, we see that (H0) is general enough to cover almost all cases that interest us (e.g., $X = C([-r, 0]; \mathbb{R}^n)$, $L^1(\mathbb{R}^n)$, etc.). Next, let U be a given metric space. For $T > 0$, the admissible control set is the following:

$$\mathcal{U}_{\text{ad}} = L^\infty(0, T; U).$$

We let

$$\Delta = \{(t, s) \in [0, T] \times [0, T] \| 0 \leqq s < t \leqq T\},$$
$$\bar{\Delta} = \{(t, s) \in [0, T] \times [0, T] \| 0 \leqq s \leqq t \leqq T\}.$$

Now, our optimal control problem can be stated as follows.

PROBLEM C. Minimize

$$J(u(\cdot)) = \int_0^T f^0(t, x(t; u), u(t))\, dt, \tag{2.1}$$

subject to the evolutionary distributed parameter system with the constraints on the initial and terminal states and the control:

$$x(t; u) = G(t, 0)x(0; u) + \int_0^t G(t, s)f(s, x(s; u), u(s))\, ds, \qquad 0 \leqq t \leqq T, \tag{2.2}$$

$$(x(0; u), x(T; u)) \in S \subset X \times X, \tag{2.3}$$

$$u(\cdot) \in \mathcal{U}_{\text{ad}}. \tag{2.4}$$

To understand our problem, let us make the following hypotheses.

(H1) The evolution operator $G: \Delta \to \mathcal{L}(Y, X)$ is strongly continuous in Δ and there exist constants $M > 0$ and $0 \leqq \alpha < 1$, such that

$$\|G(t, s)\|_{\mathcal{L}(Y,X)} \leqq \frac{M}{(t-s)^\alpha} \quad \forall (t, s) \in \Delta. \tag{2.5}$$

Moreover, the operator $G: \bar{\Delta} \to \mathcal{L}(X, X)$ is also strongly continuous and

$$G(s, s) = I \quad \forall s \in [0, T], \tag{2.6}$$

where I is the identity operator on X.

(H2) The mappings $f: [0, T] \times X \times U \to Y, f^0: [0, T] \times X \times U \to \mathbb{R}$ and their Fréchet derivatives f_x and f_x^0 are strongly continuous.

(H3) The set $S \subset X \times X$ is convex and closed.

Remark 2.1. By general end constraint, we mean that the initial and the terminal states of the system satisfy (2.3). In [25], Li and Yao considered the case of

$$S=\{x_0\}\times Q,$$

with $x_0\in X$ given and $Q\subset X$ being of finite codimension. On the other hand, Li and Chow [24] studied the optimal periodic control problems, which corresponds to the case where

$$S=\{(x,x)|x\in X\}.$$

Now, we let $\bar{u}(\cdot)$ be an optimal control and $\bar{x}(\cdot)\equiv x(\cdot\,;\bar{x}_0,\bar{u}(\cdot))$ be the corresponding optimal trajectory of the system (2.2) with the (optimal) initial state $\bar{x}_0$. We denote

$$B(t)=f_x(t,\bar{x}(t),\bar{u}(t))\quad \forall t\in[0,T], \tag{2.7}$$

and for any $y\in Y$, we define an operator $G_1(t,s)$ through the following equation:

$$G_1(t,s)y=G(t,s)y+\int_s^t G(t,r)B(r)G_1(r,s)y\,dr,\qquad 0\leqq s<t\leqq T. \tag{2.8}$$

It is not hard to show that $G_1(\cdot,\cdot)$ is well defined and that it satisfies [11]

$$G_1(t,s)y=G(t,s)y+\int_s^t G_1(t,r)B(r)G(r,s)y\,dr\quad \forall x\in X,\quad 0\leqq s<t\leqq T. \tag{2.9}$$

Moreover, for any $\eta\in X$ and $h(\cdot)\in L^1(0,T;Y)$, the solution $\xi(\cdot)$ of

$$\xi(t)=G(t,0)\eta+\int_0^t G(t,s)B(s)\xi(s)\,ds+\int_0^t G(t,s)h(s)\,ds,\qquad t\in[0,T] \tag{2.10}$$

can be written as

$$\xi(t)=G_1(t,0)\eta+\int_0^t G_1(t,s)h(s)\,ds,\qquad t\in[0,T]. \tag{2.11}$$

Next, let us denote

$$\mathcal{R}=\left\{\xi\in X|\xi=\int_0^T G_1(T,t)[f(t,\bar{x}(t),u(t))-f(t,\bar{x}(t),\bar{u}(t))]\,dt,\ u(\cdot)\in\mathcal{U}_{\text{ad}}\right\}, \tag{2.12}$$

$$Q=\{\eta\in X|\eta=x_1-G_1(T,0)x_0,(x_0,x_1)\in S\}. \tag{2.13}$$

DEFINITION 2.2. Let Ω be a subset of some topological vector space Z. We say that Ω is of finite codimension in Z, if there exists a point $z_0\in\overline{\text{co}}\,\Omega$, such that

$$\text{span}\,\{\Omega-z_0\}\equiv\text{the closed subspace spanned by }\{z-z_0|z\in\Omega\}$$

is a finite-codimensional closed subspace of Z and $\Omega-z_0$ has a nonempty interior in span $\{\Omega-z_0\}$.

It is not hard to see that if $\Omega\subset Z$ is of finite codimension in Z, then for any $z\in\overline{\text{co}}\,\Omega$, span $\{\Omega-z\}$ is a finite-codimensional closed subspace of Z. We will use this fact in sequel.

Before we state our main result, let us introduce the following Hamiltonian for our Problem C. For all $(t,x,\psi,\psi^0,u)\in[0,T]\times X\times X^*\times\mathbb{R}\times U$,

$$H(t,x,\psi,\psi^0,u)=\langle\psi,f(t,x,u)\rangle+\psi^0f^0(t,x,u). \tag{2.14}$$

Now, let us state our main result of this paper.

THEOREM 2.3 (maximum principle). *Let* (H0)-(H3) *hold. Let* $(\bar{x}(\cdot),\bar{u}(\cdot))$ *be a solution of Problem* C. *Let* $\mathcal{R}-Q$ *be of finite codimension in X. Then, there exists a* $(\psi(\cdot),\psi^0)\neq 0$, *such that*

$$\psi^0\leqq 0, \tag{2.15}$$

$$(2.16)\quad \begin{aligned}\psi(t)=G^*(T,t)\psi(T)&+\int_t^T G^*(s,t)f_x^*(s,\bar{x}(s),\bar{u}(s))\psi(s)\,ds\\ &+\psi^0\int_t^T G^*(s,t)f_x^{0*}(s,\bar{x}(s),\bar{u}(s))\,ds,\qquad t\in[0,T],\end{aligned}$$

$$(2.17)\quad H(t,\bar{x}(t),\psi(t),\psi^0,\bar{u}(t))=\max_{u\in U}H(t,\bar{x}(t),\psi(t),\psi^0,u)\quad \text{a.e. } t\in[0,T],$$

and

$$(2.18)\quad \langle\psi(0),x_0-\bar{x}(0)\rangle-\langle\psi(T),x_1-\bar{x}(T)\rangle\leqq 0\quad \forall(x_0,x_1)\in S.$$

3. Some preliminary results. In this section, we give several lemmas which will play crucial roles in the proof of the maximum principle.

LEMMA 3.1. *Let $x(\cdot)$ be the solution of* (2.2) *corresponding to $(x_0,v(\cdot))\in X\times\mathcal{U}_{\text{ad}}$. Then, for any $\rho\in(0,1]$, $\eta\in X$ and $w(\cdot)\in\mathcal{U}_{\text{ad}}$, there exists a measurable set $E_\rho\subseteq[0,T]$ with*

$$(3.1)\quad \text{meas}\,(E_\rho)=\rho T,$$

such that the solution $x_\rho(\cdot)$ of (2.2) *corresponding to $x_0+\rho\eta$ and*

$$(3.2)\quad v_\rho(t)=\begin{cases}v(t), & t\in[0,T]\backslash E_\rho,\\ w(t), & t\in E_\rho,\end{cases}$$

satisfies

$$(3.3)\quad x_\rho(t)=x(t)+\rho\delta x(t)+o(\rho)\quad \text{as } \rho\downarrow 0,$$

uniformly in $t\in[0,T]$, $w(\cdot)\in\mathcal{U}_{\text{ad}}$ and η in bounded sets of X, where $\delta x(\cdot)$ satisfies the following variational equation:

$$(3.4)\quad \begin{aligned}\delta x(t)=G(t,0)\eta&+\int_0^t G(t,s)f_x(s,x(s),v(s))\delta x(s)\,ds\\ &+\int_0^t G(t,s)[f(s,x(s),w(s))-f(s,x(s),v(s))]\,ds,\qquad t\in[0,T].\end{aligned}$$

The proof can be found in [25].

LEMMA 3.2. *Let Z be a locally convex topological vector space. Let Q be a finite-codimensional subset of Z with the property that*

$$(3.5)\quad 0\in\overline{\text{co}}\,Q.$$

Suppose $\{f_n\}$ is a sequence of continuous linear functionals on Z satisfying the following:

(i) *There exists a convex neighborhood N of zero and there exist constants c, $C>0$, such that*

$$(3.6)\quad c\leqq\sup_{z\in N}f_n(z)\leqq C\quad \forall n\geqq 1;$$

(ii) *There exists a sequence of positive numbers $\varepsilon_n\downarrow 0$, such that*

$$(3.7)\quad f_n(z)\geqq-\varepsilon_n\quad \forall n\geqq 1,\quad z\in Q;$$

(iii) *There exists a continuous linear functional f, such that*

$$(3.8)\quad f_n\to f\quad \textit{weakly}^*.$$

Then, we have

$$(3.9)\quad f\neq 0.$$

Proof. First of all, from (3.7), we see that

$$f_n(z) \geqq -\varepsilon_n \quad \forall n \geqq 1, \quad z \in \overline{\text{co}}\, Q. \tag{3.10}$$

Next, from our assumptions on Q, we can find closed subspaces Z_0 and Z_1 of Z, such that

$$Z = Z_0 \oplus Z_1, \qquad \dim Z_1 < \infty, \tag{3.11}$$

and for some $\bar{z}_0 \in \overline{\text{co}}\, Q$ and some convex neighborhood $N(\bar{z}_0)$ of $\bar{z}_0$ (in Z), it holds (note (3.5)) that

$$N(\bar{z}_0) \cap Z_0 \subset \overline{\text{co}}\, Q. \tag{3.12}$$

Without loss of generality, we may assume that the neighborhood N appearing in (i) is symmetric and bounded (in the sense that for any continuous linear functional g, we have $\sup_{z \in N} g(z) < \infty$), and

$$N(\bar{z}_0) = N + \bar{z}_0. \tag{3.13}$$

Now, suppose $f = 0$; we will obtain a contradiction. For any $z \in N$, by (3.13), we have $z + \bar{z}_0 \in N(\bar{z}_0)$. On the other hand, we have the decomposition

$$z + \bar{z}_0 = z_0 + z_1, \qquad z_0 \in Z_0, \quad z_1 \in Z_1.$$

By (3.12), we know that $z_0 \in \overline{\text{co}}\, Q$. Thus,

$$\begin{aligned} f_n(z) &= f_n(z_0) + f_n(z_1) - f_n(\bar{z}_0) \\ &\geqq -\varepsilon_n + f_n(z_1) - f_n(\bar{z}_0). \end{aligned}$$

Then, by (3.8), the assumption $f = 0$, and (3.11), we obtain

$$\delta_n \equiv |f_n(\bar{z}_0)| + |f_n(z_1)| \to 0 \quad \text{as } n \to \infty,$$

uniformly in $z \in N$ (note that z_1 depending on z). Thus, we have

$$f_n(z) \leqq \varepsilon_n + \delta_n \quad \forall z \in N, \quad n \geqq 1.$$

This leads to a contradiction to (3.6). Hence (3.9) follows. □

A result similar to Lemma 3.2 was proved by Fattorini [19] in the case where Z is a Hilbert space and by Xu [31] in the case where Z is a Banach space. Our argument is a little different from theirs and looks simpler. In § 4, we will use this result for the case where Z is a Banach space.

LEMMA 3.3. *Let Z be a Banach space and let Z_1 and Z_2 be closed subspaces of Z such that $Z = Z_1 \oplus Z_2$. Let $G_1 \in \mathscr{L}(Z)$ and set*

$$Y = \{(z, G_1 z + z_1) | z \in Z, z_1 \in Z_1\}.$$

Then, Y is a closed subspace of $Z \times Z$ and

$$Z \times Z = Y \oplus (\{0\} \times Z_2). \tag{3.14}$$

In particular, Y is of finite codimension in $Z \times Z$ if and only if Z_1 is of finite codimension in Z.

Proof. First of all, it is easy to see that Y is closed in $Z \times Z$. Next, for any $(x, y) \in Z \times Z$, there exist $z_1 \in Z_1$ and $z_2 \in Z_2$, such that

$$y - G_1 x = z_1 + z_2. \tag{3.15}$$

Thus,

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x \\ G_1 x + z_1 \end{pmatrix} + \begin{pmatrix} 0 \\ z_2 \end{pmatrix} \in Y + (\{0\} \times Z_2). \tag{3.16}$$

On the other hand, if $\binom{0}{z} \in Y \cap (\{0\} \times Z_2)$, then we have some $z_1 \in Z_1$ such that

$$\begin{pmatrix} 0 \\ z \end{pmatrix} = \begin{pmatrix} x \\ G_1 x + z_1 \end{pmatrix}.$$

Thus, $x = 0$ and

$$z = z_1 \in Z_1 \cap Z_2 = \{0\}.$$

Hence, (3.14) follows. The last conclusion is obvious. □

The next result involves the Clarke's generalized gradient. Thus, let us recall some relevant material, which can be found in [9].

Let Z be a Banach space. Let $g: Z \to \mathbb{R}$ be a Lipschitz continuous function. We define

$$g^0(z; \xi) = \overline{\lim_{z' \to z, \rho \downarrow 0}} \frac{g(z' + \rho \xi) - g(z')}{\rho} \quad \forall z, \xi \in Z, \tag{3.17}$$

$$\partial g(z) = \{\zeta \in Z^* | g^0(z; \xi) \geqq \langle \zeta, \xi \rangle, \forall \xi \in Z\} \quad \forall z \in Z. \tag{3.18}$$

It holds that

$$g^0(z; \xi) = \max \{\langle \zeta, \xi \rangle | \zeta \in \partial g(z)\} \quad \forall z, \xi \in Z. \tag{3.19}$$

Now, if Q is a convex and closed subset of Z, we can define

$$d_Q(z) = \inf_{z' \in Q} |z - z'|_Z.$$

We know that $d_Q(\cdot)$ is a convex function and

$$|d_Q(z) - d_Q(z')| \leqq |z - z'|_Z. \tag{3.20}$$

Furthermore, by Proposition 2.2.7 of [9], we have

$$\lim_{\rho \downarrow 0} \frac{d_Q(z + \rho \xi) - d_Q(z)}{\rho} = d_Q^0(z; \xi), \tag{3.21}$$

and

$$\begin{aligned} \partial d_Q(z) &\equiv \{\zeta \in Z^* | d_Q^0(z; \xi) \geqq \langle \zeta, \xi \rangle, \forall \xi \in Z\} \\ &= \{\zeta \in Z^* | d_Q(z') - d_Q(z) \geqq \langle \zeta, z' - z \rangle, \forall z' \in Z\}. \end{aligned} \tag{3.22}$$

In addition, we have the following lemma.

LEMMA 3.4. *Let Q be a convex and closed subset of some Banach space Z. Then, for any $z \notin Q$,*

$$|\zeta|_{Z^*} = 1 \quad \forall \zeta \in \partial d_Q(z). \tag{3.23}$$

Proof. By (3.20) and Proposition 2.1.2 of [9], we know that

$$|\zeta|_{Z^*} \leqq 1 \quad \forall \zeta \in \partial d_Q(z). \tag{3.24}$$

Now, since $z \notin Q$, for any $0 < \delta < 1$, there exists a $q_\delta \in Q$, such that

$$d_Q(z) \geqq (1 - \delta) |z - q_\delta|_Z > 0. \tag{3.25}$$

Thus, for any $\zeta \in \partial d_Q(z)$, by (3.22), we have

$$-d_Q(z) \geqq \langle \zeta, q_\delta - z \rangle.$$

Then,

$$(1-\delta)|z-q_\delta|_Z \leqq d_Q(z) \leqq -\langle \zeta, q_\delta - z\rangle \leqq |\zeta|_{Z^*}|z-q_\delta|_Z.$$

Hence, by (3.25),

$$|\zeta|_{Z^*} \geqq 1-\delta.$$

Then, (3.23) follows (see (3.24)). □

COROLLARY 3.5. *Let Z be a Banach space with strictly convex dual Z^*. Let Q be a convex and closed subset of Z. Then, for any $z \notin Q$, $\partial d_Q(z)$ is a singleton.*

Proof. By Proposition 2.1.2 of [9], we know that $\partial d_Q(z)$ is convex in Z^*. Thus, our assertion follows from Lemma 3.4 and the strict convexity of Z^*. □

4. Proof of the maximum principle. This section is devoted to proving the maximum principle stated in § 2. For any $(x_0, u(\cdot)) \in X \times \mathcal{U}_{ad}$, we denote the unique solution of (2.2) with $x(0)=x_0$ by $x(\cdot\,; x_0, u)$. Then, let us introduce the following

$$x^0(t; x_0, u) = \int_0^t f^0(s, x(s; x_0, u), u(s))\, ds \quad \forall t \in [0, T]. \tag{4.1}$$

It is clear that

$$x^0(T; x_0, u) = J(u(\cdot)). \tag{4.2}$$

As in § 2, we let $\bar{u}(\cdot)$ be an optimal control, $\bar{x}(\cdot)$ be the optimal trajectory of the system (2.2) corresponding to $\bar{u}(\cdot)$ and the (optimal) initial state $\bar{x}_0$ and $\bar{x}^0(\cdot)$ be the corresponding function obtained through (4.1). Then, for any $\varepsilon > 0$, we define

$$F_\varepsilon(x_0, u) = \{d_S(x_0, x(T; x_0, u))^2 + d_{S_\varepsilon^0}(x^0(T; x_0, u))^2\}^{1/2} \quad \forall (x_0, u) \in X \times \mathcal{U}_{ad}, \tag{4.3}$$

where

$$\begin{aligned} d_S(x_0, x_1) &= d((x_0, x_1), S) \\ &\equiv \inf\{(|x_0-y_0|^2 + |x_1-y_1|^2)^{1/2} | (y_0, y_1) \in S\} \quad \forall (x_0, x_1) \in X \times X, \\ S_\varepsilon^0 &= (-\infty, -\varepsilon + \bar{x}^0(T)], \\ d_{S_\varepsilon^0}(x^0) &= d(x^0, S_\varepsilon^0) \quad \forall x^0 \in \mathbb{R}. \end{aligned}$$

In the space $X \times \mathcal{U}_{ad}$, we introduce the following metric. For all $x, \hat{x} \in X$, $u(\cdot), \hat{u}(\cdot) \in \mathcal{U}_{ad}$,

$$\bar{d}((x, u(\cdot)), (\hat{x}, \hat{u}(\cdot))) = \{|x-\hat{x}|_X^2 + d(u(\cdot), \hat{u}(\cdot))^2\}^{1/2},$$

where

$$d(u(\cdot), \hat{u}(\cdot)) = \text{meas}\,\{t \in [0, T] | u(t) \neq \hat{u}(t)\}.$$

Then, it is not hard to see that $(X \times \mathcal{U}_{ad}, \bar{d})$ is a complete metric space (see [16], [19], [31]). Also, it is easy to see that $F_\varepsilon(\cdot, \cdot)$ is continuous on $X \times \mathcal{U}_{ad}$. Next, we see that

$$F_\varepsilon(x_0, u(\cdot)) > 0 \quad \forall (x_0, u(\cdot)) \in X \times \mathcal{U}_{ad}, \tag{4.4}$$

$$F_\varepsilon(\bar{x}_0, \bar{u}(\cdot)) = \varepsilon \leqq \inf F_\varepsilon(x_0, u(\cdot)) + \varepsilon. \tag{4.5}$$

Thus, by Ekeland's variational principle [16], [17], there exists a pair $(x_0^\varepsilon, u^\varepsilon(\cdot)) \in X \times \mathcal{U}_{ad}$, such that

$$\bar{d}((x_0^\varepsilon, u^\varepsilon(\cdot)), (\bar{x}_0, \bar{u}(\cdot))) \leqq \sqrt{\varepsilon}, \tag{4.6}$$

$$F_\varepsilon(x_0^\varepsilon, u^\varepsilon(\cdot)) \leqq F_\varepsilon(\bar{x}_0, \bar{u}(\cdot)), \tag{4.7}$$

$$\begin{aligned} F_\varepsilon(x_0, u(\cdot)) \geqq F_\varepsilon(x_0^\varepsilon, u^\varepsilon(\cdot)) - \sqrt{\varepsilon}\, \bar{d}((x_0^\varepsilon, u^\varepsilon(\cdot)), (x_0, u(\cdot))), \\ \forall (x_0, u(\cdot)) \in X \times \mathcal{U}_{ad}. \end{aligned} \tag{4.8}$$

Now, take any $(\eta, u(\cdot)) \in X \times \mathcal{U}_{\text{ad}}$ and $\rho \in (0, 1]$, by Lemma 3.1, we know that there exists a measurable set $E_\rho \subset [0, T]$, with meas $E_\rho = \rho T$, such that if we let

$$u_\rho^\varepsilon(t) = \begin{cases} u^\varepsilon(t), & t \in [0, T] \backslash E_\rho, \\ u(t), & t \in E_\rho, \end{cases} \tag{4.9}$$

and let $x^\varepsilon(\cdot)$, $x^{0,\varepsilon}(\cdot)$ and $x_\rho^\varepsilon(\cdot)$, $x_\rho^{0,\varepsilon}(\cdot)$ be the solutions of (2.2) and (4.1) corresponding to $(x_0^\varepsilon, u^\varepsilon(\cdot))$ and $(x_0^\varepsilon + \rho\eta, u_\rho^\varepsilon(\cdot))$, respectively; then, for any $t \in [0, T]$, as $\rho \downarrow 0$,

$$x_\rho^\varepsilon(t) = x^\varepsilon(t) + \rho\xi_\varepsilon(t) + o(\rho), \tag{4.10}$$

$$x_\rho^{0,\varepsilon}(t) = x^{0,\varepsilon}(t) + \rho\xi_\varepsilon^0(t) + o(\rho), \tag{4.11}$$

where

$$\begin{aligned} \xi_\varepsilon(t) = G(t, 0)\eta &+ \int_0^t G(t, s) f_x(s, x^\varepsilon(s), u^\varepsilon(s))\xi_\varepsilon(s)\, ds \\ &+ \int_0^t G(t, s)[f(s, x^\varepsilon(s), u(s)) - f(s, x^\varepsilon(s), u^\varepsilon(s))]\, ds, \qquad t \in [0, T], \end{aligned} \tag{4.12}$$

$$\begin{aligned} \xi_\varepsilon^0(t) = &\int_0^t f_x^0(s, x^\varepsilon(s), u^\varepsilon(s))\xi_\varepsilon(s)\, ds \\ &+ \int_0^t [f^0(s, x^\varepsilon(s), u(s)) - f^0(s, x^\varepsilon(s), u^\varepsilon(s))]\, ds, \qquad t \in [0, T]. \end{aligned} \tag{4.13}$$

Now, let us separate several cases.

Case 1. For all small enough $\varepsilon > 0$, we have

$$d_S(x_0^\varepsilon, x^\varepsilon(T)),\ d_{S_\varepsilon^0}(x^{0,\varepsilon}(T)) > 0. \tag{4.14}$$

Then, we see that for all small enough $\rho > 0$,

$$d_{S_\varepsilon^0}(x_\rho^{0,\varepsilon}(T)) > 0. \tag{4.15}$$

Thus, we have

$$d_{S_\varepsilon^0}(x_\rho^{0,\varepsilon}(T)) - d_{S_\varepsilon^0}(x^{0,\varepsilon}(T)) = \rho\xi_\varepsilon^0(T) + o(\rho). \tag{4.16}$$

On the other hand, since S is convex and closed in $X \times X$, by (3.19) and (3.21), we have

$$\begin{aligned} &\lim_{\rho\downarrow 0} \frac{d_S(x_0^\varepsilon + \rho\eta, x_\rho^\varepsilon(T)) - d_S(x_0^\varepsilon, x^\varepsilon(T))}{\rho} \\ &= d_S^0((x_0^\varepsilon, x^\varepsilon(T)); (\eta, \xi_\varepsilon(T))) \\ &= \max\{\langle a, \eta\rangle + \langle b, \xi_\varepsilon(T)\rangle | (a, b) \in \partial d_S(x_0^\varepsilon, x^\varepsilon(T))\} \\ &= \langle a_\varepsilon, \eta\rangle + \langle b_\varepsilon, \xi_\varepsilon(T)\rangle, \end{aligned} \tag{4.17}$$

where $\langle \cdot, \cdot \rangle$ is the duality between X and X^* and $\partial d_S(x_0^\varepsilon, x^\varepsilon(T))$ is the singleton $\{(a_\varepsilon, b_\varepsilon)\}$ (see Corollary 3.5). By (4.14) and Lemma 3.4, we know that

$$|a_\varepsilon|_{X^*}^2 + |b_\varepsilon|_{X^*}^2 = 1. \tag{4.18}$$

Here, we take $\{|\cdot|_X^2 + |\cdot|_X^2\}^{1/2}$ as the norm of $X \times X$. Then, in (4.8), by letting $x_0 = x_0^\varepsilon + \rho\eta$ and $u(\cdot) = u_\rho^\varepsilon(\cdot)$, we have

$$-\sqrt{\varepsilon}\,(|\eta|^2 + T^2)^{1/2} \leqq \frac{F_\varepsilon(x_0^\varepsilon + \rho\eta, u_\rho^\varepsilon(\cdot)) - F_\varepsilon(x_0^\varepsilon, u^\varepsilon(\cdot))}{\rho}. \tag{4.19}$$

Letting $\rho \to 0$ and by (4.16)-(4.17), we obtain

$$-\sqrt{\varepsilon}\,(|\eta|^2+T^2)^{1/2} \leqq \langle \bar{\varphi}_\varepsilon, \eta\rangle + \langle \bar{\psi}_\varepsilon, \xi_\varepsilon(T)\rangle + \bar{\psi}^0_\varepsilon \xi^0_\varepsilon(T), \tag{4.20}$$

where

$$\bar{\varphi}_\varepsilon = \frac{d_S(x_0^\varepsilon, x^\varepsilon(T))}{F_\varepsilon(x_0^\varepsilon, u^\varepsilon(\cdot))} a_\varepsilon, \tag{4.21}$$

$$\bar{\psi}_\varepsilon = \frac{d_S(x_0^\varepsilon, x^\varepsilon(T))}{F_\varepsilon(x_0^\varepsilon, u^\varepsilon(\cdot))} b_\varepsilon, \tag{4.22}$$

$$\bar{\psi}^0_\varepsilon = \frac{d_{S^0_\varepsilon}(x^{0,\varepsilon}(T))}{F_\varepsilon(x_0^\varepsilon, u^\varepsilon(\cdot))}. \tag{4.23}$$

Thus, by (4.18), we have

$$|\bar{\varphi}_\varepsilon|^2_{X^*} + |\bar{\psi}_\varepsilon|^2_{X^*} + (\bar{\psi}^0_\varepsilon)^2 = 1. \tag{4.24}$$

From (4.6), we know that as $\varepsilon \downarrow 0$, we have

$$\begin{gathered} x_0^\varepsilon \to \bar{x}_0, \qquad \xi_\varepsilon(t) \to \xi(t), \\ \xi^0_\varepsilon(t) \to \xi^0(t), \end{gathered} \tag{4.25}$$

uniformly in $t \in [0, T]$, where $\xi(\cdot)$ and $\xi^0(\cdot)$ satisfy the following equations:

$$\begin{aligned} \xi(t) = G(t, 0)\eta &+ \int_0^t G(t, s) f_x(s, \bar{x}(s), \bar{u}(s))\xi(s)\, ds \\ &+ \int_0^t G(t, s)[f(s, \bar{x}(s), u(s)) - f(s, \bar{x}(s), \bar{u}(s))]\, ds, \qquad t \in [0, T], \end{aligned} \tag{4.26}$$

$$\begin{aligned} \xi^0(t) = &\int_0^t f^0_x(s, \bar{x}(s), \bar{u}(s))\xi(s)\, ds \\ &+ \int_0^t [f^0(s, \bar{x}(s), u(s)) - f^0(s, \bar{x}(s), \bar{u}(s))]\, ds, \qquad t \in [0, T]. \end{aligned} \tag{4.27}$$

On the other hand, by (3.22), we have

$$\langle a_\varepsilon, x_0 - x_0^\varepsilon\rangle + \langle b_\varepsilon, x_1 - x^\varepsilon(T)\rangle \leqq -d_S(x_0^\varepsilon, x^\varepsilon(T)) \leqq 0 \quad \forall (x_0, x_1) \in S. \tag{4.28}$$

Thus,

$$\begin{aligned} &\langle \bar{\varphi}_\varepsilon, x_0 - \bar{x}_0\rangle + \langle \bar{\psi}_\varepsilon, x_1 - \bar{x}(T)\rangle \\ &\leqq (|x_0^\varepsilon - \bar{x}_0|^2_X + |x^\varepsilon(T) - \bar{x}(T)|^2_X)^{1/2} \equiv \delta_\varepsilon \to 0 \quad \text{as } \varepsilon \to 0. \end{aligned} \tag{4.29}$$

Then, by (4.20), we have

$$\begin{aligned} &\langle \bar{\varphi}_\varepsilon, \eta - (x_0 - \bar{x}_0)\rangle + \langle \bar{\psi}_\varepsilon, \xi(T) - (x_1 - \bar{x}(T))\rangle + \bar{\psi}^0_\varepsilon \xi^0(T) \\ &\geqq -\sqrt{\varepsilon}\,(|\eta|^2+T^2)^{1/2} - \delta_\varepsilon - |\xi_\varepsilon(T) - \xi(T)| - |\xi^0_\varepsilon(T) - \xi^0(T)| \\ &\geqq -\theta_\varepsilon \quad \forall (x_0, x_1) \in S. \end{aligned} \tag{4.30}$$

Here, we can see that θ_ε is uniform in $u(\cdot) \in \mathscr{U}_{\mathrm{ad}}$ and η in bounded sets of X. Now, we let

$$\hat{\mathscr{R}} = \left\{ \begin{pmatrix} \hat{\eta} \\ \hat{\xi} \end{pmatrix} \in X \times X \,\middle|\, \hat{\xi} = G_1(T, 0)\hat{\eta} + \xi,\ \xi \in \mathscr{R},\ \hat{\eta} \in X \right\},$$

and let

$$X_1 = \operatorname{span}\,(\mathcal{R} - Q + \bar{x}(T) - G_1(T, 0)\bar{x}_0).$$

Then, by our assumption, we know that X_1 is finite-codimensional in X. On the other hand, it is not hard to see that

$$\operatorname{span}\left(\hat{\mathcal{R}} - S + \begin{pmatrix} \bar{x}_0 \\ \bar{x}(T) \end{pmatrix}\right) = Y \equiv \left\{ \begin{pmatrix} x \\ G_1(T, 0)x + x_1 \end{pmatrix} \middle| x \in X, x_1 \in X_1 \right\}.$$

Thus, by Lemma 3.3, we know that Y is finite-codimensional in $X \times X$. Equivalently, we have that $\hat{\mathcal{R}} - S + \binom{\bar{x}_0}{\bar{x}(T)}$ is finite-codimensional in $X \times X$. Also, this set contains the element zero of $X \times X$. Thus, by Lemma 3.2 and (4.24), (4.30), we have a subsequence (denoted in the same way) such that

$$(\bar{\varphi}_\varepsilon, \bar{\psi}_\varepsilon, \bar{\psi}^0_\varepsilon) \xrightarrow{*} (\bar{\varphi}, \bar{\psi}, \bar{\psi}^0) \neq 0. \tag{4.31}$$

Then, from (4.20) and (4.23), we obtain

$$\langle \bar{\varphi}, \eta \rangle + \langle \bar{\psi}, \xi(T) \rangle + \bar{\psi}^0 \xi^0(T) \geqq 0, \tag{4.32}$$

$$\bar{\psi}^0 \geqq 0. \tag{4.33}$$

Case 2. There exists a sequence $\varepsilon \downarrow 0$, such that

$$d_S(x_0^\varepsilon, x^\varepsilon(T)) = 0. \tag{4.34}$$

Then, by (4.4), we must have

$$d_{S^0_\varepsilon}(x^{0,\varepsilon}(T)) > 0. \tag{4.35}$$

Thus, (4.20) and (4.21)-(4.23) read

$$-\sqrt{\varepsilon}\,(|\eta|^2 + T^2)^{1/2} \leqq \xi^0_\varepsilon(T), \tag{4.36}$$

$$\bar{\varphi}_\varepsilon = 0, \quad \bar{\psi}_\varepsilon = 0, \quad \bar{\psi}^0_\varepsilon = 1. \tag{4.37}$$

Letting $\varepsilon \to 0$, we get (trivially)

$$\xi^0(T) \geqq 0, \tag{4.38}$$

and

$$\bar{\varphi} = 0, \quad \bar{\psi} = 0, \quad \bar{\psi}^0 = 1. \tag{4.39}$$

Case 3. There exists a sequence $\varepsilon \downarrow 0$, such that

$$d_{S^0_\varepsilon}(x^{0,\varepsilon}(T)) = 0. \tag{4.40}$$

Then, by (4.4), we have

$$d_S(x_0^\varepsilon, x^\varepsilon(T)) > 0. \tag{4.41}$$

Then, by the arguments used in Cases 1 and 2, we can show that (4.32)-(4.33) hold with $\bar{\psi}^0 = 0$. Thus, in any case, we always have (4.32)-(4.33). On the other hand, from (4.28) (which is true for all of these three cases), we obtain that

$$\langle \bar{\varphi}, x_0 - \bar{x}_0 \rangle + \langle \psi, x_1 - \bar{x}(T) \rangle \leqq 0 \quad \forall (x_0, x_1) \in S. \tag{4.42}$$

Now, let $\psi(\cdot)$ be the unique solution of (2.16). Then, by (2.8) and (2.9), we have

$$\psi(t) = -G_1^*(T, t)\bar{\psi} + \psi^0 \int_t^T G_1^*(s, t) g(s)\, ds, \qquad t \in [0, T], \tag{4.43}$$

where

$$g(s)=f_x^0(s,\bar{x}(s),\bar{u}(s))^*, \qquad s\in[0,T], \tag{4.44}$$

and

$$\psi^0=-\bar{\psi}^0\leqq 0. \tag{4.45}$$

From (2.11) and (4.26)-(4.27), we know that

$$\xi(t)=G_1(t,0)\eta+\int_0^t G_1(t,s)h(s)\,ds, \qquad t\in[0,T], \tag{4.46}$$

$$\xi^0(t)=\int_0^t [\langle g(s),\xi(s)\rangle+h^0(s)]\,ds, \qquad t\in[0,T], \tag{4.47}$$

where

$$h(s)=f(s,\bar{x}(s),u(s))-f(s,\bar{x}(s),\bar{u}(s)), \qquad t\in[0,T],$$
$$h^0(s)=f^0(s,\bar{x}(s),u(s))-f^0(s,\bar{x}(s),\bar{u}(s)), \qquad t\in[0,T].$$

Thus, we have the following duality equality (note (2.14)):

$$\begin{aligned}&\langle\psi(T),\xi(T)\rangle-\langle\psi(0),\eta\rangle+\psi^0\xi^0(T)\\ &=\int_0^T [\langle\psi(t),h(t)\rangle+\psi^0h^0(t)]\,dt,\\ &=\int_0^T [H(t,\bar{x}(t),\psi(t),\psi^0,u(t))-H(t,\bar{x}(t),\psi(t),\psi^0,\bar{u}(t))]\,dt,\end{aligned} \tag{4.48}$$

for all $\eta\in X$, $u(\cdot)\in\mathscr{U}_{\mathrm{ad}}$ and $(\psi(\cdot),\psi^0,\xi(\cdot),\xi^0(\cdot))$ satisfy (4.4), (4.45)-(4.47). Hence, by setting $\eta=0$, we obtain (note (4.32))

$$\int_0^T [H(t,\bar{x}(t),\psi(t),\psi^0,u(t))-H(t,\bar{x}(t),\psi(t),\psi^0,\bar{u}(t))]\,dt\leqq 0 \quad \forall u(\cdot)\in\mathscr{U}_{\mathrm{ad}}. \tag{4.49}$$

Thus, (2.17) follows. By taking $u(\cdot)=\bar{u}(\cdot)$, we get (see (4.32) and (4.48))

$$\langle\bar{\varphi},\eta\rangle\geqq\langle\psi(0),\eta\rangle \quad \forall\eta\in X.$$

Thus, $\bar{\varphi}=\psi(0)$ and (2.18) follows. Finally, by (4.31) and (4.43), we see that $(\psi(\cdot),\psi^0)\neq 0$. The proof of the maximum principle is completed. □

Remark 4.1. At the present time it is an open question whether the strict convexity of X^* could be removed.

5. Applications. This section discusses some interesting and important cases covered by our result.

(1) The control problem with fixed endpoints. The constraint of this problem is $S=\{(x_0,x_1)\}$. Thus, we have $Q=\{x_1-G_1(T,0)x_0\}$. Hence, provided (H0)-(H3) hold and $\mathscr{R}$ is finite-codimensional in X, the maximum principle holds. This is the result of Fattorini [19].

(2) The control problem with a terminal end constraint. In this problem, we have $S=\{x_0\}\times Q_1$. Thus, $Q=Q_1-G_1(T,0)x_0$ is finite-codimensional in X is the same as Q_1 is so. Hence, provided (H0)-(H3) hold and Q_1 is finite-codimensional in X, the maximum principle holds. This is the result of Li and Yao [25]. The difference is that they did not assume the strict convexity of X^*.

(3) The control problem with separated end constraints. For this problem, $S = Q_0 \times Q_1$. We have that the maximum principle holds if (H0)-(H3) hold and $Q = Q_1 - G_1(T, 0)Q_0$ is finite-codimensional in X. This result is new.

(4) The optimal periodic control problem. The periodicity of the problem gives $S = \{(x, x) | x \in X\}$. In addition to (H0)-(H3), let us assume that $G(\cdot, \cdot)$, f^0 and f satisfy the periodic conditions:

(H4) For any $0 \leqq s \leqq t$, $x \in X$, and $u \in U$,

$$G(t+T, s+T) = G(t, s), \tag{5.1}$$

$$f(t+T, x, u) = f(t, x, u), \tag{5.2}$$

$$f^0(t+T, x, u) = f^0(t, x, u). \tag{5.3}$$

From Theorem 2.3, we have the following theorem.

THEOREM 5.1. *Let* (H0)-(H4) *hold, and let* $\mathcal{R}$ – Range $(I - G_1(T, 0))$ *be finite-codimensional in X. Then, the maximum principle holds for the periodic optimal control problem, i.e., there exist* $\psi^0 \leqq 0$ *and* $\psi(\cdot)$ *satisfying* (2.16) *with the periodic condition* $\psi(T) = \psi(0)$, *such that the maximum condition* (2.17) *holds and* $(\psi^0, \psi(\cdot)) \neq 0$.

COROLLARY 5.2. *Assume that* (H0)-(H4) *hold and* $G_1(T, 0)$ *is a compact operator on X. Then, the maximum principle holds for the optimal periodic control problem.*

Here we only need to note that when $G_1(T, 0)$ is compact, then $I - G_1(T, 0)$ is a Fredholm operator on X and thus it has a finite-codimensional range. Then, Theorem 5.1 applies.

Finally, let us give two interesting examples.

Example 5.3. Let $r > 0$, $X = C([-r, 0]; \mathbb{R}^n)$. Then, X is separable. Thus, we may endow a new norm to X so that X^* is strictly convex (see [12]-[14]). Consider the following functional differential system

$$\frac{dx(t)}{dt} = f(t, x_t, u(t)), \tag{5.4}$$

where $f: \mathbb{R} \times X \times \mathbb{R}^m \to \mathbb{R}^n$ is a given map and $x_t \in X$ is defined by

$$x_t(\theta) = x(t+\theta) \quad \forall \theta \in [-r, 0],$$

whenever $x(\cdot)$ is continuous. Furthermore, we let $f^0: \mathbb{R} \times X \times \mathbb{R}^m \to \mathbb{R}$ be given. We assume that (H2) and (5.2)-(5.3) hold for the maps f and f^0. Now, we let $G_1(\cdot, \cdot)$ be the solution operator of the variational equation

$$\frac{d\delta x(t)}{dt} = f_y(t, x_t, u(t))\delta x_t, \tag{5.5}$$

where $f_y(t, x_t, u)$ is the Fréchet derivative of $f(t, x_t, u)$ in x_t. From Hale [21], we know that $G_1(T, 0)$ is compact for $T > r$. Thus, by Corollary 5.2, if the period $T > r$, then the maximum principle holds for the periodic optimal control problem of functional differential system (5.4) without any additional condition. Here, we eliminate the conditions imposed by Colonius [10], Li [23], or Li and Chow [24] in proving the similar result. In this example, we may also take $X = W^{1,p}([-r, 0]; \mathbb{R}^n)$ with $1 \leqq p < \infty$.

Example 5.4. Let $\Omega \subset \mathbb{R}^n$ be a bounded region with smooth boundary Γ. Let $a_{ij} \in C^2(\Omega)$ such that for some constant $\alpha > 0$,

$$\sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \geqq \alpha \sum_{i=1}^{n} \xi_i^2 \quad \forall \xi = (\xi_1, \cdots, \xi_n) \in \mathbb{R}^n, \quad x \in \Omega. \tag{5.6}$$

Consider the parabolic system

$$\begin{aligned} &\frac{\partial y}{\partial t}=\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2 y}{\partial x_i\,\partial x_j}+f(t,x,y,u(x,t)), \qquad (t,x)\in\mathbb{R}\times\Omega,\\ &y(T,x)=y(0,x), \qquad x\in\Omega,\\ &y(t,x)=0, \qquad (t,x)\in\mathbb{R}\times\Gamma, \end{aligned} \tag{5.7}$$

where $f:\mathbb{R}\times\Omega\times\mathbb{R}\times\mathbb{R}^m\to\mathbb{R}$ is a continuous function such that f_y is continuous and

$$f(t+T,\cdot,\cdot,\cdot)=f(t\cdot,\cdot,\cdot).$$

If we let

$$Ay=-\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2 y}{\partial x_i\,\partial x_j},$$

with the Dirichlet boundary condition, then e^{-At} is an analytic semigroup on $X\equiv L^2(\Omega)$ which is compact for all $t>0$ [20], [27]. Now, let

$$J(u(\cdot))=\int_0^T\int_\Omega f^0(t,x,y(t,x),u(t,x))\,dx\,dt. \tag{5.8}$$

We consider the following periodic optimal control problem.

Minimize $J(\cdot)$ subject to the parabolic system (5.7), with

$$\begin{aligned} &u(t,\cdot)\in\mathcal{U}_{\mathrm{ad}}\subseteq L^\infty(\Omega) \quad \text{a.e. } t\in[0,T],\\ &u(t+T,\cdot)=u(t,\cdot). \end{aligned} \tag{5.9}$$

Now, if $(\bar y(\cdot,\cdot),\bar u(\cdot,\cdot))$ is optimal, then it is not hard to show that the evolution operator $G_1(t,s)$ of the variational equation:

$$\begin{aligned} &\frac{\partial(\delta y)}{\partial t}=\sum_{i,j=1}^{n} a_{ij}(x)\frac{\partial^2(\delta y)}{\partial x_i\,\partial x_j}+f_y(t,x,\bar y(t,x),\bar u(x,t))\delta y, \qquad (t,x)\in\mathbb{R}\times\Omega,\\ &y(0,x)=y_0(x), \qquad x\in\Omega,\\ &y(t,x)=0, \qquad (t,x)\in\mathbb{R}\times\Gamma, \end{aligned} \tag{5.10}$$

has the property that $G_1(T,0)$ is compact. Thus, the maximum principle holds for this periodic optimal control problem.

Acknowledgments. The authors thank the referees for pointing out a mistake in the original version of the paper.

REFERENCES

[1] N. U. AHMED AND K. L. TEO, *Optimal Control of Distributed Parameter Systems*, North-Holland, Amsterdam, 1981.

[2] E. ASPLUND, *Averaged norms*, Israel J. Math., 5 (1967), pp. 227-233.

[3] ———, *Fréchet-differentiability of convex functions*, Acta Math., 121 (1968), pp. 31-47.

[4] A. V. BALAKRISHNAN, *Applied Functional Analysis*, Springer-Verlag, New York, 1976.

[5] H. T. BANKS AND G. KENT, *Control of functional equations to target sets in function space*, SIAM J. Control Optim., 10 (1972), pp. 567-593.

[6] L. D. BERKOVITZ, *Optimal Control Theory*, Springer-Verlag, New York, 1974.

[7] A. G. BUTKOVSKII, *Maximum principle of optimal control for distributed parameter systems*, Avtomat. Telemekh., 22 (1961), pp. 1288-1301. (In Russian.)

[8] F. H. CLARKE, *The maximum principle with minimum hypotheses*, SIAM J. Control Optim., 14 (1976), pp. 1078-1091.

[9] ———, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[10] F. COLONIUS, *Optimality for periodic control of functional differential systems*, Report #36, Math. Inst. University of Graz, Graz, Austria, 1984.

[11] R. F. CURTAIN AND A. J. PRITCHARD, *Infinite Dimensional Linear Systems Theory*, Lecture Notes in Control and Information Sciences, Vol. 8, Springer-Verlag, New York, 1981.

[12] M. M. DAY, *Strict convexity and smoothness of normed spaces*, Trans. Amer. Math. Soc., 78 (1955), pp. 516-528.

[13] ———, *Normed Linear Spaces*, 3rd ed., Springer-Verlag, Berlin, New York, 1973.

[14] K. DEIMLING, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, New York, 1985.

[15] J. DIESTEL, *Geometry of Banach Spaces*, Lecture Notes in Mathematics, Vol. 485, Springer-Verlag, Berlin, New York, 1975.

[16] I. EKELAND, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324-353.

[17] ———, *Nonconvex minimization problems*, Bull Amer. Math. Soc. (NS), 1 (1979), pp. 443-474.

[18] H. O. FATTORINI, *The maximum principle for nonlinear nonconvex systems in infinite dimensional spaces*, Lecture Notes in Control and Information Sciences, Vol. 75, Springer-Verlag, New York, 1985.

[19] ———, *A unified theory of necessary conditions for nonlinear nonconvex control systems*, Appl. Math. Optim., 15 (1987), pp. 141-185.

[20] A. FRIEDMAN, *Partial Differential Equations*, Holt, Reinhart and Winston, New York, 1969.

[21] J. K. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[22] F. L. JIN AND X. J. LI, *On optimal control of functional differential systems, Distributed Parameter Systems*, Lecture Notes in Control and Information Sciences, Vol. 102, Springer-Verlag, New York, 1987, pp. 112-119.

[23] X. J. LI, *Maximum principle of optimal periodic control for functional differential systems*, J. Optim. Theory Appl., 50 (1986), pp. 421-429.

[24] X. J. LI AND S. N. CHOW, *Maximum principle of optimal control for functional differential systems*, J. Optim. Theory Appl., 54 (1987), pp. 335-360.

[25] X. J. LI AND Y. YAO, *Maximum principle of distributed parameter systems with time lags*, Distributed Parameter Systems, Lecture Notes in Control and Information Sciences, Vol. 75, Springer-Verlag, New York, 1985, pp. 410-427.

[26] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[27] A. PAZY, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[28] L. S. PONTRYAGIN, *The maximum principle in the theory of optimal processes*, Proc. 1st Congress IFAC, Moscow, 1960.

[29] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISCHENKO, *Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

[30] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

[31] H. XU, *Necessary conditions of optimal control for distributed parameter systems*, M.S. thesis, Fudan University, Shanghai, China, 1988.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 909-925, July 1991

010

OPTIMAL CONTROL WITH INFINITE HORIZON FOR DISTRIBUTED PARAMETER SYSTEMS WITH CONSTRAINED CONTROLS*

G. DI BLASIO†

Abstract. Optimal control problems for systems governed by linear partial differential state equations with constrained controls are studied. If the dynamics are stabilizable with respect to the cost, it is proved that the value function is a generalized solution for the associated stationary Hamilton-Jacobi equation. Moreover, the feedback formula and uniqueness are proved under suitable assumptions.

Key words. Hamilton-Jacobi equations, optimal control problem, constrained controls

AMS(MOS) subject classifications. 49C20, 34G20

1. Introduction and statement of results. Let X and U be real separable Hilbert spaces with inner product $\langle\,,\rangle_X$ and $\langle\,,\rangle_U$, respectively. We will be concerned with the following problem. Minimize the functional

$$J_\infty(u,x)=\int_0^{+\infty}[h(u(t))+g(y(t))]\,dt \tag{1}$$

over all measurable $u: R_+\to U$ and $y: R_+\to X$ satisfying the state equation (in the mild sense)

$$y'(t)=Ay(t)+Bu(t),\quad t\geqq 0,\qquad y(0)=x \tag{2}$$

and the control constraint

$$u(t)\in C,\qquad t\geqq 0. \tag{3}$$

Here

(i) $h: U\to R$ is defined as $h(u)=\frac{1}{2}\|u\|_U^2$ if $u\in C$, whereas $h(u)=+\infty$ if $u\notin C$;

(ii) $g: X\to R_+$ is a lower semicontinuous (l.s.c.) convex function;

(iii) $A: D(A)\subseteq X\to X$ is the infinitesimal generator of a C_0-semigroup $S(t)$ satisfying $\|S(t)x\|_X\leqq \exp(\omega t)\|x\|_X$, for some $\omega\in R$;

(iv) $B\in L(U,X)$;

(v) $C\subseteq U$ is a closed, convex, and bounded set containing the origin.

We will follow the dynamic programming approach, which exploits the connection of (1)-(3) with the following stationary Hamilton-Jacobi equation:

$$H(B^*\varphi_x(x))-\langle\varphi_x(x),Ax\rangle_X=g(x), \tag{4}$$

where B^* is the adjoint of B and $H: U\to R$ is defined as

$$H(u)=\tfrac{1}{2}[\|u\|_U^2-\|u+P_C(-u)\|_U^2]$$

(here $P_C(u)$ denotes the projection of u on C).

In the case where $g(\cdot)=\frac{1}{2}\|\cdot\|_X^2$ and $C=U$ problem (1)-(3) reduces to the well-known linear-quadratic optimal control problem and (4) is replaced by the algebraic Riccati equation. For this problem and the connection with the algebraic Riccati equation there is an almost complete theory (see, e.g., the review paper of Pritchard and Zabczyk [12] and the references therein).

* Received by the editors December 27, 1989; accepted for publication (in revised form) August 10, 1990. This work was partially supported by Ministero Pubblica Istruzione, and by Consiglio Nazionale Ricerche.

† Dipartimento di Matematica, Università di Roma I, Piazzale Aldo Moro 5, 00185 Rome, Italy.

In this paper we prove that there exists an optimal pair for (1)-(3) if and only if (4) admits a solution. Moreover, we will give sufficient conditions for the existence of solutions of (4). These solutions will be obtained as

$$\varphi(x) = \lim_{t\to\infty} \underline{\varphi}(t, x),$$

where $\underline{\varphi}$ is the solution of the Hamilton-Jacobi equation

$$\varphi_t(t, x) + H(B^*\varphi_x(t, x)) - \langle \varphi_x(t, x), Ax\rangle_X = g(x) \tag{5}$$

satisfying

$$\varphi(0, x) = 0.$$

Stationary Hamilton-Jacobi equations have been studied by Crandall and Lions [5]-[7] in the case where the equations contain an additional term $\varepsilon\varphi$ with $\varepsilon > 0$. This regularization does not allow for the study of problem (1)-(3). We also recall the results of Cannarsa and Da Prato [4] who have studied (1)-(3) and the corresponding Hamilton-Jacobi equation in the case where h and g are locally Lipschitz and convex functions and the operator A is replaced by $A_1 + F$, where A_1 generates an analytic semigroup and F is nonlinear.

The plan of the paper is as follows. In § 2 we recall some results concerning an approximated version of (5). In §§ 3-4 we prove existence and establish some properties of solutions of (5). Furthermore, in § 5 we study the existence of solutions of (4) together with the connection with problem (1)-(3). Moreover, we give sufficient conditions for the feedback formula and for uniqueness. Finally in § 6 we consider an application to a partial differential state equation.

2. Preliminaries. The approximating Hamilton–Jacobi equation. In this section we recall some results of [8], which will be needed in the sequel. We use the following notation for functions $\varphi: X \to R$:

$$C_2(X) = \{\varphi \text{ is continuous}, \sup\,(|\varphi(x)|/(1+\|x\|^2)) < +\infty\},$$

$$|\varphi(\cdot)|_R = \sup_{\|x\|\leqq R} |\varphi(x)|,$$

$$C^1_{\mathrm{Lip}}(X) = \{\varphi \text{ is Fréchet-differentiable}, \sup_{x\neq y}((\|\varphi'(x)-\varphi'(y)\|)/(\|x-y\|)) < +\infty\}$$

(here and in the following the subscripts X and U in the scalar products are omitted for simplicity in notation).

Let $F: X \to R$ be the function defined as $F(x) = H(B^*x)$

$$F(x) = \tfrac{1}{2}(\|B^*x\|^2 - \|B^*x + P_C(-B^*x)\|^2); \tag{6}$$

then it can be proved (see [8, § 2]) that F is convex, Fréchet differentiable and

$$F'(x) = -BP_C(-B^*x). \tag{7}$$

Furthermore, we denote by F^* the conjugate function of F

$$F^*(x) = \sup_{y\in X}\{\langle x, y\rangle - F(y)\}. \tag{8}$$

Hence F^* satisfies the property

$$\langle x, y\rangle \leqq F^*(x) + F(y) \quad \text{with equality holding iff } x = F'(y). \tag{9}$$

Now let K_0 be the subset of $C^1_{\mathrm{Lip}}(X)$ defined as

$$K_0 = \{\varphi \in C^1_{\mathrm{Lip}}(X): \varphi \text{ is convex}, \varphi(0) = \varphi'(0) = 0\}. \tag{10}$$

For given $\varphi \in K_0$ and $\lambda > 0$ we set

$$\varphi_\lambda(x) = \inf_{y \in X} \left\{ \varphi(y) + \lambda F^*\left(\frac{x-y}{\lambda}\right)\right\}. \tag{11}$$

It can be proved that the infimum is attained at $y = J_\lambda(x)$, where

$$J_\lambda(x) = (1 + \lambda F'\varphi')^{-1}(x). \tag{12}$$

Therefore

$$\varphi_\lambda(x) = \varphi(J_\lambda(x)) + \lambda F^*\left(\frac{x - J_\lambda(x)}{\lambda}\right). \tag{13}$$

In the case where $F(x) = \frac{1}{2}\|x\|^2$, (i.e., when $B = I$ and $C = U$), the functions φ_λ given by (11) reduce to the well-known regularization of convex functions. In the present context it can be proved that φ_λ satisfy the following properties (see [8, § 3]):

(i) $\varphi_\lambda \in K_0$, $\varphi_\lambda(x) \leqq \varphi(x)$,

(ii) $\lim_{\lambda \downarrow 0} \varphi_\lambda(x) = \varphi(x)$,

(iii) $\lim_{\lambda \downarrow 0} (\varphi(x) - \varphi_\lambda(x))/\lambda = F(\varphi'(x))$.

In view of (iii) we will replace the Hamilton-Jacobi equation (5) by the following approximating equation:

$$\varphi_t(t, x) + 1/\lambda(\varphi(t, x) - \varphi_\lambda(t, x)) - \langle Ax, \varphi_x(t, x)\rangle = g(x). \tag{14}$$

Equation (14) and the initial condition

$$\varphi(0, x) = f(x) \tag{15}$$

yield to the integral equation

$$\begin{aligned} \varphi(t, x) = \exp\left(-\frac{t}{\lambda}\right) f(S(t)x) + \int_0^t \exp\left(\frac{s-t}{\lambda}\right) \\ \cdot \left[\frac{1}{\lambda}\varphi_\lambda(s, S(t-s)x) + g(S(t-s)x)\right] ds, \end{aligned} \tag{16}$$

where $S(t)$ is the semigroup generated by A. Conversely, if φ satisfy (16), then it is easy to see that for each $x \in D(A)$ the function $t \to \varphi(t, x)$ is continuously differentiable and satisfies (14)-(15).

We have the following lemma (see [8, Thm. 3]).

LEMMA 1. *Let $f, g \in K_0$. Then there exists a unique $\varphi^\lambda \in C(0, T; C_2(X))$ such that $\varphi^\lambda(t, \cdot) \in K_0$, $t \to \varphi(t, x)$ is continuously differentiable for $x \in D(A)$, and φ^λ is a solution of* (16). *Moreover, for each $R > 0$ we have*

$$|\varphi^\lambda(t, \cdot)|_R \leqq |f(\cdot)|_{R\exp(\omega t)} + t|g(\cdot)|_{R\exp(\omega t)}, \tag{17}$$

where ω satisfies $\|S(t)\| \leqq \exp(\omega t)$.

Proof. The existence of a unique solution of (16) is proved in Theorem 3 of [7]. To prove (17) set

$$\psi(t, x) = \varphi^\lambda(t, \exp(-\omega t)x). \tag{18}$$

Then ψ satisfies

$$\psi_t(t, x) + \frac{1}{\lambda}[\psi(t, x) - (\varphi^\lambda)_\lambda(t, \exp(-\omega t)x)] - \langle A_1 x, \psi_x(t, x)\rangle = g(\exp(-\omega t)x),$$

where $A_1 = A - \omega I$. Therefore

$$\psi(t, x) = \exp\left(-\frac{t}{\lambda}\right) f(S_1(t)x) + \int_0^t \exp\left(\frac{s-t}{\lambda}\right)$$
$$\cdot \left[\frac{1}{\lambda}(\varphi^\lambda)_\lambda(s, S_1(t-s)\exp(-\omega s)x) + g(S_1(t-s)\exp(-\omega s)x)\right] ds,$$

where S_1 is the semigroup generated by A_1. Now from property (i) we have

$$(\varphi^\lambda)_\lambda(s, S_1(t-s)\exp(-\omega s)x) \leqq \varphi^\lambda(s, S_1(t-s)\exp(-\omega s)x) = \psi(s, S_1(t-s)x)$$

so that

$$|\psi(t, \cdot)|_R \leqq \exp\left(-\frac{t}{\lambda}\right)|f(\cdot)|_R + \int_0^t \exp((s-t)/\lambda)\left[\frac{1}{\lambda}|\psi(s, \cdot)|_R + |g(\cdot)|_R\right] ds$$

and hence

$$\exp\left(\frac{t}{\lambda}\right)|\psi(t, \cdot)|_R \leqq |f(\cdot)|_R + \int_0^t \exp\left(\frac{s}{\lambda}\right)|g(\cdot)|_R\, ds + \int_0^t \exp\left(\frac{s}{\lambda}\right)\frac{1}{\lambda}|\psi(s, \cdot)|_R\, ds.$$

Therefore from the Gronwall inequality we get

$$\exp\left(\frac{t}{\lambda}\right)|\psi(t, \cdot)|_R \leqq |f(\cdot)|_R + \lambda\left[\exp\left(\frac{t}{\lambda}\right) - 1\right]|g(\cdot)|_R$$
$$+ \int_0^t \exp\left(\frac{t-s}{\lambda}\right)\frac{1}{\lambda}\left[|f(\cdot)|_R + \lambda\left(\exp\left(\frac{s}{\lambda}\right) - 1\right)|g(\cdot)|_R\right] ds$$

so that (17) follows from (18). □

The following lemma establishes a monotonicity result for the solutions of (16).

LEMMA 2. *Let φ and $\tilde\varphi$ be the solutions of* (16) *with data* (f, g) *and* $(\tilde f, \tilde g)$, *respectively. If $\tilde f \leqq f$ and $\tilde g \leqq g$, then $\tilde\varphi \leqq \varphi$.*

Proof. The existence of solutions of (16) can be proved by a fixed point argument (see [8, Thm. 3]). Therefore the assertion $\tilde\varphi \leqq \varphi$ is a consequence of the fact that $\tilde\varphi \leqq \varphi$ imply $\tilde\varphi_\lambda \leqq \varphi_\lambda$, where φ_λ is defined by (11). □

3. Solutions to the Hamilton–Jacobi equation.

DEFINITION 1. Let $f, g \in K_0$ (where K_0 is defined by (10)). We say that a continuous function $\varphi: [0, T] \times X \to R$ is a *strong solution* of the problem

$$\begin{aligned} &\varphi_t(t, x) + H(B^*\varphi_x(t, x)) - \langle Ax, \varphi_x(t, x)\rangle = g(x), \\ &\varphi(0, x) = f(x), \end{aligned} \tag{19}$$

if

(i) $\varphi(t, \cdot) \in K_0$; $\varphi(\cdot, x) \in \text{Lip}(0, T)$, for each $x \in D(A)$,

(ii) there exist $\{f_n\}, \{g_n\} \subseteq K_0$ and $\{\varphi_n\} \subseteq C(0, T; C_2(X))$ verifying $\varphi_n(t, \cdot) \in K_0$; for each $x \in D(A)$ $t \to \varphi_n(t, x)$ is continuously differentiable. Moreover,

$$f_n \to f, \quad \text{in } C_2(X), \quad g_n \to g \quad \text{in } C_2(X), \quad \varphi_n \to \varphi \quad \text{in } C(0, T; C_2(X)),$$

and

$$\frac{\partial}{\partial t}\varphi_n(t, x) + H(B^*\varphi_{nx}(t, x)) - \langle Ax, \varphi_{nx}(t, x)\rangle = g_n(x),$$

$$\varphi_n(0, x) = f(x).$$

We have the following existence result.

THEOREM 1. *For each $f, g \in K_0$ there exists a unique strong solution of problem* (19) *and we have*

(i) $\varphi = \lim_{\lambda \downarrow 0} \varphi^\lambda$ *in* $C(0, T; C_2(X))$, *where* φ^λ *is the solution of* (16).

Moreover, we have

(ii) $|\varphi(t, \cdot)|_R \leqq |f(\cdot)|_{R\exp(\omega t)} + t|g(\cdot)|_{R\exp(\omega t)}$.

Furthermore, if φ and $\tilde{\varphi}$ are strong solutions of (19) *with data (f, g) and $(\tilde{f}, \tilde{g})$, respectively, and $\tilde{f} \leqq f$, $\tilde{g} \leqq g$, then*

(iii) $\tilde{\varphi}(t, \cdot) \leqq \varphi(t, \cdot)$.

Proof. Assertion (i) is proved in Theorem 4 of [8]. Assertions (ii) and (iii) are consequences of property (i), (17), and Lemma 2. □

Now set

$$J(u, x) = f(y(T)) + \int_t^T [h(u(s)) + g(y(s))]\, ds \tag{20}$$

and consider the problem of minimizing $J(u, x)$ over all (y, u) satisfying

$$y'(s) = Ay(s) + Bu(s), \quad t \leqq s \leqq T, \qquad y(t) = x. \tag{21}$$

The following theorem concerns the connection between strong solutions of (19) and problems (20)-(21).

THEOREM 2. *Let φ be the strong solution of* (19). *Then for each (y, u) satisfying* (21) (*in the mild sense*) *we have*

$$\varphi(T-t, x) + \int_t^T \frac{1}{2} \|P_C(-B^*\varphi_x(T-s, y(s))) - u(s)\|^2\, ds \leqq J(u, x). \tag{22}$$

If in addition (y^*, u^*) in (21) *satisfies*

$$u^*(s) = P_C(-B^*\varphi_x(T-s, y^*(s)),$$

then

$$\varphi(T-t, x) = f(y^*(T)) + \int_t^T [h(u^*(s)) + g(y^*(s))]\, ds. \tag{23}$$

Proof. Using (31) of [7], we have

$$\varphi(T-t, x) + \int_t^T [H(B^*\varphi_x(T-s, y(s))) + \langle \varphi_x(T-s, y(s)), Bu(s)\rangle + \tfrac{1}{2}\|u\|^2]\, ds \leqq J(u, x).$$

Furthermore, from the convexity of C we have

$$\langle B^*\varphi_x(T-s, y(s)) + P_C(-B^*\varphi_x(T-s, y(s))), u(s) - P_C(-B^*\varphi_x(T-s, y(s)))\rangle \geqq 0$$

so that from (6)

$$\begin{aligned} &H(B^*\varphi_x(T-s, y(s))) + \langle \varphi_x(T-s, y(s)), Bu(s)\rangle \\ &\quad = -\tfrac{1}{2}\|P_C(-B^*\varphi_x(T-s, y(s)))\|^2 - \langle B^*\varphi_x(T-s, y(s)), P_C(-B^*\varphi_x(T-s, y(s)))\rangle \\ &\qquad + \langle B^*\varphi_x(T-s, y(s)), u(s)\rangle \\ &\quad \geqq \tfrac{1}{2}\|P_C(-B^*\varphi_x(T-s, y(s)))\|^2 - \langle P_C(-B^*\varphi_x(T-s, y(s))), u(s)\rangle \end{aligned}$$

and (22) is proved. Assertion (23) is proved in Theorem 5 of [8]. □

In the following we need existence for solutions of (19) with more general data (f, g). To this end we introduce the subset K_0'' defined as

$$K_0'' = \{\varphi : X \to R\colon \varphi \text{ is convex and l.s.c., } 0 \in \operatorname{Int} D(\varphi),\ \varphi(0) = 0, 0 \in \partial\varphi(0)\}. \tag{24}$$

As usual, we denote by $D(\varphi)$ the effective domain of φ

$$D(\varphi)=\{x\in X\colon \varphi(x)<+\infty\}$$

and by $\partial\varphi$ the subdifferential of φ

$$\partial\varphi(x)=\{x'\in X\colon \varphi(x+y)\geqq \varphi(x)+\langle y, x'\rangle, \forall y\in X\}.$$

DEFINITION 2. Let $f, g\in K_0''$. We say that a function φ from $[0, T]\times X$ into R is a *generalized solution* of (19) if

(i) $\varphi(t,\cdot)\in K_0''$,

(ii) there exist $\{f_\varepsilon\}, \{g_\varepsilon\}\subseteq K_0$ and $\{\varphi_\varepsilon\}$ such that

$$f_\varepsilon\uparrow f,\quad g_\varepsilon\uparrow g,\quad \varphi_\varepsilon(t,x)\to\varphi(t,x),$$

moreover, φ_ε is the strong solution of the problem

$$\begin{aligned}&\varphi_t(t,x)+H(B^*\varphi_x(t,x))-\langle Ax, \varphi_x(t,x)\rangle=g_\varepsilon,\\ &\varphi(0,x)=f_\varepsilon(x).\end{aligned}\tag{25}$$

We have the following existence result.

THEOREM 3. *For each $f, g\in K_0''$ there exists a generalized solution of* (19).

Proof. For each $\varepsilon>0$ set

$$\begin{aligned}f_\varepsilon(x)&=\min_{y\in X}\left[\frac{1}{2\varepsilon}\|x-y\|^2+f(y)\right],\\ g_\varepsilon(x)&=\min_{y\in X}\left[\frac{1}{2\varepsilon}\|x-y\|^2+g(y)\right].\end{aligned}\tag{26}$$

We have $f_\varepsilon, g_\varepsilon\in K_0$ and moreover

$$f_\varepsilon(x)\uparrow f(x),\qquad g_\varepsilon(x)\uparrow g(x).\tag{27}$$

Using Theorem 1(ii) and (iii) and (27) we have that there exists φ_ε strong solution of (25) and moreover

$$|\varphi_\varepsilon(t,\cdot)|_R\leqq|f_\varepsilon(\cdot)|_{R\exp(\omega t)}+t|g_\varepsilon(\cdot)|_{R\exp(\omega t)}\tag{28}$$

and

$$\varphi_\varepsilon(t,x)\leqq\varphi_{\varepsilon'}(t,x),\qquad 0<\varepsilon'<\varepsilon.$$

Therefore there exists φ such that

$$\varphi_\varepsilon(t,x)\uparrow\varphi(t,x).$$

Moreover, we have that $\varphi(t,\cdot)$ is convex and l.s.c. To accomplish the proof it suffices to show that $0\in\operatorname{Int}D(\varphi(t,\cdot))$. Since $0\in\operatorname{Int}D(f)$, $\operatorname{Int}D(g)$ and f and g are convex we have that there exists R' such that $B(0,R')\subseteq D(f), D(g)$ (here $B(0,r)=\{x\in X\colon \|x\|\leqq r\}$). Therefore using (27) and (28), we have that $B(0, R'/\exp(\omega T))\subseteq D(\varphi(t,\cdot))$ so that $0\subseteq\operatorname{Int}D(\varphi(t,\cdot))$. □

We now investigate the connection between the generalized solutions of (19) and the optimal control problem (20)-(21). To this end we set for (y, u) satisfying (21)

$$J_\varepsilon(u,x)=f_\varepsilon(y(T))+\int_t^T[h(u(s))+g_\varepsilon(y(s))]\,ds.$$

We have the following theorem.

THEOREM 4. *Let φ be a generalized solution of* (19). *Then we have*

$$\varphi(T-t,x)=\inf J(u,x).\tag{29}$$

Moreover, if $x \in D(\varphi(T-t, \cdot))$ *then there exists the optimal pair* (y^*, u^*) *for* $J(u, x)$ *and we have*

$$y^* = w - \lim y_\varepsilon, \qquad u^* = w - \lim u_\varepsilon, \tag{30}$$

where $(y_\varepsilon, u_\varepsilon)$ *is optimal for* $J_\varepsilon(u, x)$.

Proof. Let $\{\varphi_\varepsilon\}$, $\{f_\varepsilon\}$, and $\{g_\varepsilon\}$ be given by Definition 2(ii). By (22) we have

$$\varphi_\varepsilon(T-t, x) + \int_t^T \frac{1}{2} \|P_C(-B^*\varphi_{\varepsilon,x}(T-s, y(s))) - u(s)\|^2 \, ds \leqq J_\varepsilon(u, x) \leqq J(u, x). \tag{31}$$

Therefore letting $\varepsilon \to 0$ we get

$$\varphi(T-t, x) \leqq J(u, x). \tag{32}$$

To accomplish the proof of (29) let $(y_\varepsilon, u_\varepsilon)$ in (21) be given by

$$u_\varepsilon(s) = P_C(-B^*\varphi_{\varepsilon,x}(T-s, y_\varepsilon(s))).$$

By (23) we have

$$\varphi_\varepsilon(T-t, x) = J_\varepsilon(u, x). \tag{33}$$

Moreover, since $u_\varepsilon \in C$ and C is bounded we have, for $s \in [t, T]$, $\|u_\varepsilon(s)\| \leqq \text{const}$, so that $\|y_\varepsilon(s)\| \leqq \text{const}$. Therefore there exists a subsequence $\varepsilon_k \to 0$ satisfying, for almost every $s \in [t, T]$,

$$u_{\varepsilon_k}(s) \to u^*(s), \qquad y_{\varepsilon_k}(s) \to y^*(s), \quad \text{weakly} \tag{34}$$

and moreover (y^*, u^*) satisfies (21) (in the mild sense). Now let $\varepsilon' > 0$; since $\{f_\varepsilon\}$ is monotone we have for $\varepsilon_k < \varepsilon'$

$$f_{\varepsilon_k}(y_{\varepsilon_k}(T)) \geqq f_{\varepsilon'}(y_{\varepsilon_k}(T))$$

so that

$$\liminf f_{\varepsilon_k}(y_{\varepsilon_k}(T)) \geqq f_{\varepsilon'}(y^*(T)),$$

which in turn implies

$$\liminf f_{\varepsilon_k}(y_{\varepsilon_k}(T)) \geqq f(y^*(T)).$$

In the same way we find

$$\liminf g_{\varepsilon_k}(y_{\varepsilon_k}(s)) \geqq g(y^*(s)).$$

Hence we get from (33)

$$\varphi(T-t, x) \geqq J(u^*, x) \tag{35}$$

so that from (32)

$$\varphi(T-t, x) = J(u^*, x)$$

and (29) is proved. Moreover, if $x \in D(\varphi(T-t, \cdot))$, then (y^*, u^*) is optimal. Since the optimal pair is unique, by the same argument used above we prove that each subsequence of $(y_\varepsilon, u_\varepsilon)$ contains a subsequence that converges weakly to (y^*, u^*). □

4. Properties of the solutions of the Hamilton–Jacobi equation. In this section we establish a number of properties of the solutions of (19). We begin with the following result.

THEOREM 5. *For each $f, g \in K_0''$ there exists a unique generalized solution of* (19). *If, in addition, $D(f) = D(g) = X$, then we have*

(i) $D(\varphi(t, \cdot)) = X$,

(ii) *$\varphi(t, \cdot)$ is locally Lipschitz continuous on X, uniformly for $t \in [0, T]$,*

(iii) *$\varphi(\cdot, x)$ is Lipschitz continuous on $[0, T]$ for each $x \in D(A)$,*

(iv) *φ is continuous on $[0, T] \times X$,*

(v) *the optimal pair (y^*, u^*) satisfies the feedback formula $u^*(s) \in P_C(-B^* \partial_x \varphi(T - s, y^*(s)))$, for a.e. $s \in [0, T]$ (where $\partial_x \varphi(t, x)$ denotes the subdifferential of $x \to \varphi(t, x)$).*

Proof. The existence and the uniqueness of φ are proved in Theorem 3 and in (29). To prove (i) and (ii) let us use (29) with $u = 0$

$$\varphi(T - t, x) \leqq f(\exp(A(T - t))x + \int_t^T g(\exp(A(s - t))x)\, ds.$$

Since $D(f) = D(g) = X$ we have, by well-known properties of l.s.c. and convex functions, that f and g are continuous on X. Therefore the functions $F_1(x) = \sup_{s \in [0, T]} f(\exp(As)x)$ and $G_1(x) = \sup_{s \in [0, T]} g(\exp(As)x)$ are finite for each $x \in X$. Since they are l.s.c. and convex functions, we have that F_1 and G_1 are continuous on X. Thus for each $x \in X$ there exist r_x and M_x such that

$$\varphi(T - t, y) \leqq M_x, \quad y \in B(x, 2r_x), \quad t \in [0, T] \tag{36}$$

and (i) is proved. Moreover, we have (see, e.g., [9, Cor. 1.2.4]) that $\varphi(t, \cdot)$ is Lipschitz continuous on $B(x, r_x)$ with Lipschitz constant $M_x r_x^{-1}$, which proves (ii). To prove (iii) fix $t_1, t_2 \in [0, T]$ with $t_1 < t_2$ and let (y_1^*, u_1^*) be optimal at (t_1, x), we have

$$\begin{aligned} |\varphi(T - t_1, x) - \varphi(T - t_2, x)| &\leqq |\varphi(T - t_1, x) - \varphi(T - t_2, y_1^*(t_2))| \\ &\quad + |\varphi(T - t_2, y_1^*(t_2)) - \varphi(T - t_2, x)| = I_1 + I_2. \end{aligned}$$

Since the restriction of (y_1^*, u_1^*) to $[t_2, T]$ is optimal at $(t_2, y_1^*(t_2))$, we have

$$I_1 = \int_{t_1}^{t_2} h(u_1^*(s)) + g(y_1^*(s))\, ds = I_1' + I_1''.$$

Since $u_1^*(s) \in C$, and C is bounded, we have $\|u_1^*(s)\| \leqq$ const. so that

$$I_1' \leqq c(t_2 - t_1).$$

Moreover, if $x \in D(A)$

$$\begin{aligned} \|y_1^*(s) - x\| &\leqq \|\exp((s - t_1)A)x - x\| + \left\| \int_{t_1}^{s} Bu_1^*(\sigma)\, d\sigma \right\| \\ &\leqq \left\| \int_0^{s - t_1} A \exp(\sigma A)x\, d\sigma \right\| + c(t_2 - t_1) \\ &\leqq (\exp(\omega T)\|Ax\| + c)(t_2 - t_1). \end{aligned}$$

Since g is continuous at x there exist r_x and M_x such that if $y \in B(x, r_x)$, then $g(y) \leqq M_x$. Therefore if $t_2 - t_1 \leqq r_x' = r_x(\exp(\omega T)\|Ax\| + c) - 1$ we have

$$I_1'' \leqq M_x(t_2 - t_1)$$

so that if $t_2 - t_1 \leqq r_x'$

$$I_1 \leqq I_1' + I_1'' \leqq c_1(x)(t_2 - t_1).$$

Furthermore, using (ii) we have that there exist r''_x and $c_2(x)$ such that for $t_2-t_1 \leqq r''_x$ we get

$$I_2 \leqq c_2(x)(t_2-t_1),$$

which achieves the proof of (iii). Moreover, (iv) follows from (ii) and (36). To prove (v) let us use (31) with (y, u) replaced by (y^*, u^*)

$$\varphi_\varepsilon(T-t, x)+\int_t^T \frac{1}{2}\|P_C(-B^*\varphi_{\varepsilon,x}(T-s, y^*(s)))-u^*(s)\|^2\, ds \leqq J(u, x).$$

Since (y^*, u^*) is optimal we have $\varphi(T-t, x)=J(u^*, x)$; hence there exists $\{\varepsilon_k\}$ such that, for almost every $s\in[0, T]$,

$$\lim P_C(-B^*\varphi_{\varepsilon_k,x}(T-s, y^*(s)))=u^*(s). \tag{37}$$

Now fix $t_0\in[0, T]$. From (36) there exist r_0 and M_0 such that if $s\in[0, T]$ and $\|x-y^*(t_0)\|\leqq r_0$, then $\varphi(T-s, x)\leqq M_0$, so that $\varphi_\varepsilon(T-s, x)\leqq M_0$; moreover, there exists σ_0 such that $|s-t_0|\leqq\sigma_0$ implies $|y^*(s)-y^*(t_0)|\leqq r_0/4$. Therefore using the abovementioned property of convex functions (see [9]) we have that for $s\in[t_0-\sigma_0, t_0+\sigma_0]$ the function $\varphi_\varepsilon(T-s, \cdot)$ is Lipschitz continuous on the ball $B(y^*(s), r_0/4)$ and the Lipschitz constant $c=c(M_0, r_0)$ is independent of ε. Therefore for $s\in[t_0-\sigma_0, t_0-\sigma_0]$ we have

$$\|\varphi_{\varepsilon,x}(T-s, y^*(s))\|\leqq c(M_0, r_0).$$

Since $[0, T]$ is compact we have that there exists c such that for $s\in[0, T]$ we have

$$\|\varphi_{\varepsilon,x}(T-s, y^*(s))\|\leqq c;$$

therefore there exists a subsequence (again denoted by $\{\varepsilon_k\}$) such that

$$w-\lim \varphi_{\varepsilon_k,x}(T-s, y^*(s))=z(s)\in\partial\varphi_x(T-s, y^*(s))$$

so that by (37)

$$u^*(s)=P_C(-B^*z(s))$$

and (v) is proved. □

Remark 1. Due to Asplund's theorem [1] (see also [10, Thm. 2.12]) we have that if $D(f)=D(g)=X$, then the generalized solution of (19) is Fréchet differentiable on a dense subset of X.

The following result concerns monotonicity of the solutions of (19).

THEOREM 6. *Let φ and $\tilde{\varphi}$ be the generalized solutions of* (19) *with data* (f, g) *and* $(\tilde{f}, \tilde{g})$, *respectively. If $\tilde{f}\leqq f$ and $\tilde{g}\leqq g$, then*

$$\tilde{\varphi}(t, x)\leqq\varphi(t, x).$$

Proof. From Theorem 3 and from uniqueness we have that $\varphi=\lim\varphi_\varepsilon$, where φ_ε is the strong solution of (25) with data $(f_\varepsilon, g_\varepsilon)$ given by (26), and similarly for $\tilde{\varphi}$. Hence the assertion follows from Theorem 1(iii) and from the fact that $\tilde{f}_\varepsilon\leqq f_\varepsilon$ and $\tilde{g}_\varepsilon\leqq g_\varepsilon$. □

The following result concerns continuous dependence upon the data.

THEOREM 7. *Let φ and φ_n be the generalized solutions of* (19) *with data* (f, g) *and* (f_n, g_n), *respectively. If $f_n\uparrow f$ and $g_n\uparrow g$, then*

$$\varphi_n(t, x)\uparrow\varphi(t, x).$$

Proof. From Theorem 6 we have that $\{\varphi_n\}$ is nondecreasing. Therefore for each (t, x) there exists

$$\psi(t, x) = \lim \varphi_n(t, x). \tag{38}$$

Furthermore, from (29) we have

$$\varphi_n(T-t, x) \leqq f_n(y(T)) + \int_t^T [h(u(s)) + g_n(y(s))]\, ds \tag{39}$$

over all (y, u) satisfying (21), whereas we have

$$\varphi_n(T-t, x) = f_n(y_n(T)) + \int_t^T [h(u_n(s)) + g_n(y_n(s))]\, ds \tag{40}$$

if $x \in D(\varphi_n(T-t, \cdot))$ and (y_n, u_n) is optimal. Now (38) and (39) imply

$$\psi(T-t, x) \leqq f(y(T)) + \int_t^T [h(u(s)) + g(y(s))]\, ds$$

and hence if φ denotes the generalized solution of (19)

$$\psi(t, x) \leqq \varphi(t, x).$$

Moreover, let $x \in D(\psi(T-t, \cdot))$; using (40) and a computation similar to the one used in proving (34) we get that there exists a subsequence (y_{n_k}, u_{n_k}) such that $y_{n_k} \to y^*$ and $u_{n_k} \to u^*$, weakly, and that

$$\psi(T-t, x) \geqq f(y^*(T)) + \int_t^T [h(u^*(s)) + g(y^*(s))]\, ds.$$

Therefore we have

$$\psi(T-t, x) = \inf J(u, x), \tag{41}$$

where the infimum is taken over all (y, u) satisfying (21). Hence using (29) and (41) we get $\varphi = \psi$ and the assertion is proved. □

Remark 2. Without essential modifications it can be proved that the assertion of Theorem 7 remains true if the assumptions $f_n \uparrow f$ and $g_n \uparrow g$ are replaced by $f_n \downarrow f$ and $g_n \downarrow g$.

The following lemmas are a consequence of Theorem 6.

LEMMA 3. *Let* $\underline{\varphi}$ *be the generalized solution of problem* (19) *with* $f = 0$. *Then* $\underline{\varphi}(\cdot, x)$ *is nondecreasing, for each* $x \in X$.

Proof. For each $t_0 > 0$ we have that $\underline{\varphi}(t + t_0, x)$ is the generalized solution of (19) with $f(x) = \underline{\varphi}(t_0, x) \geqq 0$. Therefore using Theorem 6 we get $\underline{\varphi}(t, x) \leqq \underline{\varphi}(t + t_0, x)$ and the assertion is proved. □

LEMMA 4. *Let* φ *be the generalized solution of* (19), *and let* $\underline{\varphi}$ *be the generalized solution of* (19) *with* $f = 0$. *Then we have*

$$\underline{\varphi}(t, x) \leqq \varphi(t, x).$$

Proof. The assertion is a consequence of Theorem 6. □

We conclude this section with the following definition which is suggested by Lemma 4.

DEFINITION 3. We say that $\underline{\varphi}$ is the minimal solution of

$$\varphi_t(t, x) + H(B^* \varphi_x(t, x)) - \langle Ax, \varphi_x(t, x) \rangle = g(x) \tag{42}$$

if $\underline{\varphi}$ is the generalized solution of (19) with $f = 0$.

5. The stationary Hamilton–Jacobi equation and the infinite horizon problem.

DEFINITION 4. We say that a function $\varphi \in K_0''$ is a *stationary solution* of equation (42) if φ is a generalized solution of (19) with $f(x) = \varphi(x)$.

A stationary solution of (42) will be also called a solution of the equation

$$H(B^*\varphi_x(x)) + \langle Ax, \varphi_x(x)\rangle = g(x). \tag{43}$$

Moreover (43) will be called the *stationary* Hamilton-Jacobi *equation.* Equation (43) generalizes the well-known algebraic Riccati equation. For this equation uniqueness is false in general. Hence we do not expect to have uniqueness for solutions of (43). Therefore it is convenient to introduce the following notation.

DEFINITION 5. We say that $\underline{\varphi}_\infty$ is the *minimal solution* of the stationary Hamilton-Jacobi equation (43) if $\underline{\varphi}_\infty$ is a solution of (43) and for every solution φ of (43) we have

$$\underline{\varphi}_\infty(x) \leqq \varphi(x).$$

The *maximal solution* $\bar{\varphi}_\infty$ is similarly defined.

The following theorem gives sufficient and necessary conditions for the existence of the minimal solution of (43).

THEOREM 8. *Let* $\underline{\varphi}(t, x)$ *be the minimal solution of* (42). *Then* (43) *admits a minimal solution if and only if there exist* $r > 0$ *and* M_x *such that for each* $\|x\| \leqq r$

$$\underline{\varphi}(t, x) \leqq M_x. \tag{44}$$

Moreover, the minimal solution $\underline{\varphi}_\infty$ *is given by*

$$\underline{\varphi}_\infty(x) = \lim_{t\to+\infty} \underline{\varphi}(t, x). \tag{45}$$

Furthermore, we have

$$\underline{\varphi}_\infty(x) = \inf\left\{\underline{\varphi}_\infty(y(T)) + \int_0^T [h(u(s)) + g(y(s))]\, ds\right\}, \tag{46}$$

where the infimum is taken over all (y, u) *satisfying* (21) *with* $t = 0$.

Proof. By Lemma 3 we have that $\underline{\varphi}(\cdot, x)$ is nondecreasing, so that there exists $\underline{\varphi}_\infty$ given by (45). Now let φ be a solution of (43); from Lemma 4 we have $\underline{\varphi}(t, x) \leqq \varphi(x)$ so that $\underline{\varphi}_\infty \leqq \varphi$. Therefore if (43) has a solution, then (44) is satisfied. Conversely, let (44) hold; then we have that $0 \in \operatorname{Int} D(\underline{\varphi}_\infty)$ so that $\underline{\varphi}_\infty \in K_0''$. Furthermore, set $f_n(x) = \underline{\varphi}(n, x)$ and denote by φ_n the generalized solution of (19) with $f = f_n$. Then we have $f_n \uparrow \underline{\varphi}_\infty$ and $\varphi_n(t, x) \uparrow \underline{\varphi}_\infty(x)$ so that by Theorem 7 the function $\underline{\varphi}_\infty$ is a stationary solution of (42). Finally, assertion (46) is a consequence of (29). □

The following theorem establishes a monotonicity result for the minimal solutions of (43).

THEOREM 9. *Let* $\underline{\varphi}_\infty$ *and* $\tilde{\underline{\varphi}}_\infty$ *be the minimal solutions of* (43) *with data g and* $\tilde{g}$, *respectively, and let* $\tilde{g} \leqq g$. *Then we have*

$$\tilde{\underline{\varphi}}_\infty(x) \leqq \underline{\varphi}_\infty(x).$$

Proof. From Lemma 4 we have that $\tilde{\underline{\varphi}}(t, x) \leqq \underline{\varphi}(t, x)$ so that the assertion follows by a passage to the limit. □

Furthermore, the following result concerns continuous dependence upon the data.

THEOREM 10. *Let* $\underline{\varphi}_\infty$ *and* $\underline{\varphi}_{n,\infty}$ *be the minimal solutions of* (43) *with data g and* g_n, *respectively. If* $g_n \uparrow g$, *then*

$$\underline{\varphi}_{n,\infty} \uparrow \underline{\varphi}.$$

Proof. By Theorem 9 we have that $\{\varphi_{n,\infty}\}$ is nondecreasing so that there exists $\psi = \lim \varphi_{n,\infty}$. Moreover, from Theorem 7, ψ is a solution of (43). Furthermore, from Theorem 9 we have $\psi \leqq \underline{\varphi}_\infty$ so that $\psi = \underline{\varphi}_\infty$. □

We now investigate the connection between the minimal solution of (43) and the infinite horizon problem introduced in § 1.

Set

$$J_\infty(u, x) = \int_0^{+\infty} [h(u(s)) + g(y(s))]\, ds,$$

where (y, u) satisfies

$$y'(t) = Ay(t) + Bu(t), \quad t \geqq 0, \qquad y(0) = x. \tag{47}$$

DEFINITION 6. Given $x \in X$ we say that a control $u \in L^2(R_+; U)$ is *admissible* if $J_\infty(u, x) < +\infty$. Moreover, we say that $x \in X$ is *admissible* if there exists an admissible control at x. We denote the set of all admissible initial data by X_0. Moreover, if u is admissible and y verifies (47), we say that (y, u) is an *admissible pair.*

DEFINITION 7. We say that (A, B) is *locally C-stabilizable* with respect to the observation g if $0 \in \operatorname{Int} X_0$. We say that (A, B) is *C-stabilizable* (with respect to g) if $X_0 = X$.

The notion of stabilizability introduced in Definition 7 is a generalization of the one given in the linear quadratic optimal control problem (i.e., in the case where $C = U$ and $g(\cdot) = \frac{1}{2}\|\cdot\|^2$). In this case the stabilizability problem and the connection with the algebraic Riccati equation (which replaces the stationary Hamilton–Jacobi equation (43)) is widely investigated (see, e.g., the review paper [12] and the references therein).

To our knowledge there is no general theory concerning C-stabilizability. In the following we give some sufficient conditions which may be applied to some situations.

We assume that there exist c, γ, and r such that if $\|x\| \leqq r$, then $g(x) \leqq c\|x\|^\gamma$.

Example 1. The simplest example of stabilizability is the case where A is exponentially stable, i.e.,

$$\|\exp(tA)\| \leqq \exp(\omega t) \tag{48}$$

with $\omega < 0$. In this case the control $u = 0$ is admissible for each $\|x\| \leqq r$ so that $B(0, r) \subseteq X_0$. If in addition $D(g) = X$ (i.e., $g(x) < +\infty$ for each x), then $X_0 = X$ and (A, B) is C-stabilizable for each B and C.

Example 2. If the control constraint C is given by

$$C = \{u \in U: \|u\| \leqq K\}, \tag{49}$$

then if A satisfies (48) with $\omega = 0$ and $B^{-1} \in L(X, U)$, we have that (A, B) is locally C-stabilizable. In this case for fixed $x \in B(0, r)$ it suffices to take

$$u(t) = -\varepsilon B^{-1} \exp[t(A - \varepsilon I)]x,$$

where I is the identity operator, and

$$y(t) = \exp[t(A - \varepsilon I)]x,$$

where $\varepsilon \leqq K(\|B^{-1}\| r)^{-1}$. Therefore we have $B(0, r) \subseteq X_0$; moreover, if $D(g) = X$ we have $X_0 = X$ and (A, B) is C-stabilizable.

Example 3. Finally, another example of a C-stabilizable pair (A, B) (with C given by (49)) is the case where (A, B) is stabilizable with respect to I in the sense of linear quadratic optimal control problem. In this case there exists $L \in L(X, U)$ such that

$A-BL$ is exponentially stable, i.e., there exists constants M, $\alpha>0$ verifying $\|\exp[t(A-BL)]\|\leqq M\exp(-\alpha t)$. Now define $r^*=\min(rM^{-1}, KM^{-1}\|L\|^{-1})$; then for each $x\in B(0, r^*)$ it suffices to take

$$u(t)=L\exp[t(A-BL)]x, \qquad y(t)=\exp[t(A-BL)]x$$

so that we have $B(0, r^*)\subseteq X_0$.

We now investigate the connection between the stabilizability of (A, B) and the existence of solutions of (43).

THEOREM 11. *Let there exist the minimal solution of* (43). *Then* (A, B) *is locally C-stabilizable with respect to g and we have* $D(\underline{\varphi}_\infty)=X_0$.

Moreover, for each $x\in X_0$ *there exists an optimal pair* (y^*, u^*) *and we have*

$$\underline{\varphi}_\infty(x)=J_\infty(u^*, x). \tag{50}$$

If in addition $X_0=X$ *and* $D(g)=X$, *then*

$$u^*(t)\in P_C(-B^*\,\partial\underline{\varphi}_\infty(y^*(t))). \tag{51}$$

Proof. Let $\underline{\varphi}$ be the minimal solution of (42); by (29) we have

$$\underline{\varphi}(T, x)=\inf\left\{\int_0^T [h(u(s))+g(y(s))]\,ds\right\} \tag{52}$$

over all (y, u) satisfying

$$y'(s)=Ay(s)+Bu(s), \quad 0\leqq s\leqq T, \qquad y(0)=x. \tag{53}$$

Now let $x\in D(\underline{\varphi}_\infty)$ and let (y_n, u_n) be optimal on $[0, n]$; we have

$$\underline{\varphi}(n, x)=\int_0^n [h(u_n(s))+g(y_n(s))]\,ds. \tag{54}$$

Moreover, let u_n^* be defined as

$$u_n^*(s)=\begin{cases} u_n(s) & \text{if } 0\leqq s\leqq n,\\ 0 & \text{if } n<s\end{cases}$$

and let y_n^* be the solution of (53) with u replaced by u_n^*. Since $\underline{\varphi}(n, x)\leqq\underline{\varphi}_\infty(x)$ we have by (54) that $\{u_n^*\}$ is bounded on $L^2(R_+; U)$. Hence there exists $\{u_{n_k}^*\}$ such that

$$u_{n_k}^*\to u^* \quad \text{weakly in } L^2(R_+; U).$$

Therefore for each $T>0$ we have

$$y_{n_k}^*\to y^* \quad \text{weakly in } L^2(0, T; X).$$

Moreover y^* is the solution of (53) with u replaced by u^*. Furthermore, for $n_k>T$ we have

$$\underline{\varphi}_\infty(x)\geqq\underline{\varphi}(n_k, x)\geqq\int_0^T [h(u_{n_k}^*(s))+g(y_{n_k}^*(s))]\,ds$$

so that

$$\underline{\varphi}_\infty(x)\geqq\int_0^T [h(u(s^*))+g(y^*(s))]\,ds.$$

Since T is arbitrary we get

$$\underline{\varphi}_\infty(x)\geqq\int_0^{+\infty} [h(u^*(s))+g(y^*s)]\,ds \tag{55}$$

so that (y^*, u^*) is admissible and $x \in X_0$. Hence (A, B) is locally C-stabilizable and $D(\underline{\varphi}_\infty) \subseteq X_0$. Conversely if $x \in X_0$ and (y, u) is admissible at x, we have by (52)

$$\underline{\varphi}(T, x) \leqq \int_0^T [h(u(s)) + g(y(s))]\, ds \leqq J_\infty(u, x) \tag{56}$$

so that $x \in D(\underline{\varphi}_\infty)$ and hence $X_0 = D(\underline{\varphi}_\infty)$. Moreover, from (55) and (56) we have

$$\underline{\varphi}_\infty(x) = J_\infty(u^*, x).$$

Finally, (51) follows from (v) of Theorem 5. □

COROLLARY 1. *The following properties are equivalent*:

(i) (A, B) *is locally C-stabilizable with respect to* g;

(ii) *There exists a solution of* (43).

Proof. (I)⇒(ii) Let $x \in X_0$ and let (y, u) be an admissible pair at x. If $\underline{\varphi}$ is the minimal solution of (41) we have, by (29),

$$\underline{\varphi}(T, x) \leqq \int_0^T [h(u(s)) + g(y(s))]\, ds \leqq J_\infty(u, x)$$

so that the assertion follows from Theorem 8. Assertion (ii)⇒(i) is proved in Theorem 11. □

We now give sufficient conditions for the existence of the maximal solution of (43). We begin with the following result.

THEOREM 12. *Let g be coercive, that is, $g(x) \geqq \gamma\|x\|^p$, for some $p \geqq 1$. Moreover, let (A, B) be locally C-stabilizable with respect to g. Then there exists the maximal solution $\bar{\varphi}_\infty$ of* (43) *and we have $\bar{\varphi}_\infty = \underline{\varphi}_\infty$, that is, there exists a unique solution of* (43).

Proof. Let (y^*, u^*) be optimal for $J_\infty(u, x)$ and let φ be a solution of (43). Since φ is a solution of (42) with initial datum φ we have by (29)

$$\varphi(x) \leqq \varphi(y^*(t)) + \int_0^t [h(u^*(s)) + g(y^*(s))]\, ds,$$

hence by (50)

$$\varphi(x) \leqq \varphi(y^*(t)) + \int_0^{+\infty} [h(u^*(s)) + g(y^*(s))]\, ds = \varphi(y^*(t)) + \underline{\varphi}_\infty(x). \tag{57}$$

Since g is coercive we have that $y^* \in L^p(R_+; X)$ for some $p \geqq 1$. This, in turn, implies that there exists $\{t_n\} \uparrow +\infty$ such that $y^*(t_n) \to 0$. Therefore by a passage to the limit we get $\varphi(x) \leqq \underline{\varphi}_\infty(x)$, so that $\underline{\varphi}_\infty$ is maximal. □

Now let us introduce the following notation:

$$g_{(\varepsilon)}(x) = g(x) + \tfrac{1}{2}\varepsilon\|x\|^p,$$

$$J_{\varepsilon,\infty}(u, x) = \int_0^{+\infty} [h(u(s)) + g_{(\varepsilon)}(y(s))]\, ds.$$

We have the following existence result.

THEOREM 13. *Assume that (A, B) is locally C-stabilizable with respect to $g_{(1)}$, for some $p \geqq 1$. Then there exists the maximal solution of* (43) *and we have*

$$\varphi_{\varepsilon,\infty} \downarrow \bar{\varphi}_\infty, \tag{58}$$

where $\varphi_{\varepsilon,\infty}$ is the (*unique*) *solution of*

$$H(B^*\varphi_x(x)) + \langle Ax, \varphi_x(x)\rangle = g_{(\varepsilon)}(x).$$

Proof. By Theorem 9 we have that $\{\varphi_{\varepsilon,\infty}\}$ is nonincreasing so that there exists $\bar{\varphi}_\infty$ given by (58). Moreover, from Remark 2 it follows that $\bar{\varphi}_\infty$ is a solution of (43). To prove that $\bar{\varphi}_\infty$ is maximal let φ be a solution of (43) and let $(y_\varepsilon^*, u_\varepsilon^*)$ be optimal for $J_{\varepsilon,\infty}$. Using (57) with g replaced by $g_{(\varepsilon)}$ we get, since $g_{(\varepsilon)}$ is coercive, $\varphi(x) \leqq \varphi_{\varepsilon,\infty}(x)$. Therefore the conclusion follows by a passage to the limit. □

The following result concerns the connection between the maximal solution of (43) and the infinite horizon problem. We have the following theorem.

THEOREM 14. *Let (A, B) be locally C-stabilizable with respect to $g_{(1)}$ and let $\bar{\varphi}_\infty$ be the maximal solution of* (43). *Moreover, set*

$$Y = \{(y, u) \colon y \in L^2(R_+; X) \text{ and } u \in L^2(R_+; U)\}.$$

Then we have

$$\bar{\varphi}_\infty(x) = \inf J_\infty(u, x),$$

where the infimum is taken over all $(y, u) \in Y$ satisfying (47).

Proof. Let $(y, u) \in Y$ and let $\varphi_{\varepsilon,\infty}$ be defined as in the proof of Theorem 13. By (50) we have

$$\varphi_{\varepsilon,\infty} \leqq J_{\varepsilon,\infty}(u, x)$$

so that by (58) we get $\bar{\varphi}_\infty \leqq J_\infty(u, x)$ and hence $\bar{\varphi}_\infty \leqq \inf_Y J_\infty(u, x)$. Furthermore, let $(y_\varepsilon, u_\varepsilon)$ be optimal for $J_{\varepsilon,\infty}$; we have

$$\varphi_{\varepsilon,\infty}(x) = J_{\varepsilon,\infty}(u_\varepsilon, x) \geqq J_\infty(u_\varepsilon, x) \geqq \inf_Y J_\infty(u, x)$$

and the conclusion follows by a passage to the limit.

The following result concerns monotonicity of the maximal solutions of (43).

THEOREM 15. *Assume that (A, B) is locally C-stabilizable with respect to $g_{(1)}$. Moreover, let $\tilde{g} \leqq g$ and let $\tilde{\bar{\varphi}}_\infty$ and $\bar{\varphi}_\infty$ be the maximal solutions of* (43) *with data $\tilde{g}$ and g, respectively. Then we have*

$$\tilde{\bar{\varphi}}_\infty \leqq \bar{\varphi}_\infty.$$

Proof. The result follows from (58) and from Theorem 9. □

Finally, the following result concerns continuous dependence upon the data.

THEOREM 16. *Let g and g_n be such that $g_n \downarrow g$, and let (A, B) be locally C-stabilizable with respect to $g_{1,(1)}$. Moreover, let φ and $\bar{\varphi}_n$ be the maximal solutions of* (43) *with data g and g_n, respectively. Then $\bar{\varphi}_n \downarrow \bar{\varphi}$.*

Proof. The result follows from (58) and from Theorem 9. □

6. Applications.

In this section we give some examples of optimal control problems which can be studied by means of the results of § 5.

Let Ω be an open bounded set of R^n with smooth boundary $\partial\Omega$ and set $X = L^2(\Omega)$. We denote by A the operator defined as

$$D(A) = \left\{ y \in W^{2,2}(\Omega) \colon \frac{\partial y}{\partial \nu} = 0 \right\},$$

$$Ay = \Delta y, \tag{59}$$

where Δ is the Laplacian operator. Then it is known that A generates a semigroup $\exp(tA)$ verifying

$$\|\exp(tA)y\|_X \leqq \|y\|_X. \tag{60}$$

Moreover, $\sigma(A)$ consists of isolated nonpositive eigenvalues. We denote by X_1 and X_2 the subspaces corresponding to the spectral sets $\sigma_1=\{0\}$ and $\sigma_2=\sigma(A)\backslash\{0\}$. Then

$$X=X_1\oplus X_2 \tag{61}$$

and X_1, X_2 are invariant under A. Moreover, if A_i is the restriction of A to X_i, then $\exp(tA_2)$ is exponentially stable, i.e.,

$$\|\exp(tA_2)y\|_X \leqq M\exp(\omega t)\|y\|_X \tag{62}$$

for some $\omega<0$. Now let U be a real separable Hilbert space and let $B\in L(U,X)$. If $B^{-1}\in L(X,U)$ using (60) and the results of Example 2 of § 5 we have that (A,B) is C-stabilizable with respect to $\frac{1}{2}\|\cdot\|^2$ for any C of the form (49). Therefore we get the following theorem.

THEOREM 17. *Consider the problem of minimizing the functional*

$$J_\infty(u,y_0)=\frac{1}{2}\int_0^{+\infty}dt\int_\Omega [|y(t,x)|^2+|u(t,x)|^2]\,dx \tag{63}$$

over all (y,u) *verifying*

$$\begin{aligned} &y_t(t,x)=\Delta y(t,x)+(Bu)(t,x), \qquad t>0,\quad x\in\Omega,\\ &\frac{\partial}{\partial v}y(t,x)=0, \qquad t>0,\quad x\in\partial\Omega,\\ &y(0,x)=y_0(x), \qquad x\in\Omega,\\ &\int_\Omega |u(t,x)|^2\,dx\leqq K, \qquad t>0, \end{aligned} \tag{64}$$

where $B\in L(X)$. *Then if* $B^{-1}\in L(X)$ *we have*

(i) *For each* $y_0\in L^2(\Omega)$ *there exists an optimal pair* (y^*,u^*) *at* y_0 *for problem* (63), (64). *Moreover we have* $\varphi_\infty(x)=J_\infty(u^*,y_0)$, *where* φ_∞ *is the solution of*

$$\tfrac{1}{2}(\|B^*\,\partial\varphi(y)\|^2-\|B^*\,\partial\varphi(y)+P_C(-B^*\,\partial\varphi(y))\|^2)-\langle\Delta y,\partial\varphi(y)\rangle=\tfrac{1}{2}\|y\|^2$$

and

$$P_C(y)=\begin{cases} y & \text{if } \|y\|\leqq K,\\ \|y\|/K & \text{if } \|y\|>K.\end{cases}$$

Moreover the following feedback formula holds:

(ii) $u^*(t,\cdot)\in P_C(-B^*\,\partial\varphi_\infty(y^*(t,\cdot)))$.

Proof. The result is a consequence of Theorems 11 and 12 and Corollary 1. □

Now let $\psi\in L_\infty(\Omega)$ and let $B\in L(R,X)$ be given by

$$(Bu)(x)=u\cdot\psi(x). \tag{65}$$

We consider the following problem. Minimize the functional

$$J_\infty(u,y_0)=\frac{1}{2}\int_0^{+\infty}dt\left[\int_\Omega |y(t,x)|^2\,dx+|u(t)|^2\right] \tag{66}$$

over all (y,u) satisfying

$$\begin{aligned} &y_t(t,x)=\Delta y(t,x)+u(t)\psi(x), \qquad t>0,\quad x\in\Omega,\\ &\frac{\partial}{\partial \nu}y(t,x)=0, \qquad t>0,\quad x\in\partial\Omega,\\ &y(0,x)=y_0(x), \qquad x\in\Omega,\\ &|u(t)|\leqq K. \end{aligned} \tag{67}$$

To study problem (66), (67) we use the results of Example 3 of § 5 and look for conditions guaranteeing stabilizability in the sense of linear quadratic optimal control (with respect to I) for the pair (A, B) given by (59) and (65). To this end we use (61) and set $\psi=\psi_1+\psi_2$, with $\psi_i \in X_i$; moreover, we denote by $B_i \in L(R, X_i)$ the operator given by

$$(B_i u)(x)=u\psi_i(x). \tag{68}$$

By (62) we have that (A_2, B_2) is stabilizable. Therefore using a result of [11] we have that if (A_1, B_1) is stabilizable then so is (A, B). Finally, it is easy to see that (A_1, B_1) is stabilizable if and only if $\psi_1 \neq 0$, i.e.,

$$\int_\Omega \psi(x)\,dx \neq 0. \tag{69}$$

Therefore using the results of Example 3, Theorems 11 and 12 and Corollary 1, we get the following theorem.

THEOREM 18. *Let ψ satisfy* (69). *Then for each $y_0 \in L^2(\Omega)$ such that $\|y_0\|$ is suitably small there exists an optimal pair (y^*, u^*) at y_0 for problem* (66), (67), *and we have*

$$J_\infty(u^*, y_0)=\varphi_\infty(y_0),$$

where φ_∞ is the minimal solution of

$$\tfrac{1}{2}(\langle\psi, \partial\varphi(y)\rangle)^2-|\langle\psi, \partial\varphi(y)\rangle+P_C(-\langle\psi, \partial\varphi(y)\rangle)|^2-\langle\Delta y, \partial\varphi(y)\rangle=\tfrac{1}{2}\|y\|^2$$

and

$$P_C(u)=\begin{cases} u & \text{if } |u| \leqq K, \\ |u|/K & \text{if } |u|>K. \end{cases}$$

Acknowledgments. The author thanks the referees for helpful comments.

REFERENCES

[1] E. ASPLUND, *Fréchet differentiability of convex functions*, Acta Math., 121 (1968), pp. 31-47.

[2] V. BARBU AND G. DA PRATO, *Hamilton-Jacobi Equations in Hilbert Spaces*, Pitman, Boston, MA, 1982.

[3] ———, *Hamilton-Jacobi equations in Hilbert spaces. Variational and semigroup approaches*, Ann. Mat. Pura Appl., 142 (1985), pp. 303-349.

[4] P. CANNARSA AND G. DA PRATO, *Nonlinear optimal control with infinite horizon for distributed parameter systems and stationary Hamilton-Jacobi equations*, SIAM J. Control Optim., 27 (1989), pp. 861-875.

[5] M. G. CRANDALL AND P. L. LIONS, *Hamilton-Jacobi in infinite dimensions. Uniqueness of viscosity solutions*, J. Funct. Anal., 62 (1985), pp. 379-396.

[6] ———, *Hamilton-Jacobi in infinite dimensions. Part* II. *Existence of viscosity solutions*, J. Funct. Anal., 65 (1986), pp. 368-405.

[7] ———, *Hamilton-Jacobi in infinite dimension. Part* III, J. Funct. Anal., 68 (1986), pp. 214-247.

[8] G. DI BLASIO, *Global solutions for a class of Hamilton-Jacobi equations in Hilbert spaces*, Numer. Funct. Anal. Optim., 8 (1985), pp. 261-300.

[9] I. EKELAND AND R. TEMAM, *Convex Analysis and Variational Problems*, Studies in Mathematics and Its Applications, North-Holland, Amsterdam, 1976.

[10] R. R. PHELPS, *Convex Functions, Monotone Operators and Differentiability*, Lecture Notes in Mathematics, Vol. 1364, Springer-Verlag, Berlin, 1989.

[11] A. J. PRITCHARD AND R. TRIGGIANI, *On the stabilizability problem in Banach space*, J. Math. Anal. Appl., 52 (1975), pp. 383-403.

[12] A. J. PRITCHARD AND J. ZABCZYK, *Stability and stabilizability of infinite-dimensional systems*, SIAM Rev., 23 (1981), pp. 25-52.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 926-935, July 1991

011

ON OUTPUT FEEDBACK VIA GRASSMANNIANS*

XIAOCHANG WANG†

Abstract. The output feedback pole placement map χ has been studied by using the central projection model. Necessary and sufficient conditions for χ being dominant are given. 2×2 systems of McMillan degree 4 and 2×3 systems of McMillan degree 6 are studied intensively.

Key words. pole placement of linear multivariable systems, Grassmannians, central projection

AMS(MOS) subject classification. 93D15

1. Introduction. Consider a linear system

$$\dot{x} = Ax + Bu, \qquad y = Cx,$$

where x, u, y, are n, m, p, vectors, respectively. The poles of the system are the eigenvalues of A. If an output feedback

$$u = Ky$$

is applied to the system, the closed-loop system becomes

$$\dot{x} = (A + BKC)x.$$

The closed-loop poles became the eigenvalues of $A+BKC$. Identifying a characteristic polynomial with a point in affine n-space $\mathbf{A}^n$, we can define pole placement map $\chi : \mathbf{A}^{mp} \rightarrow \mathbf{A}^n$ by

$$\chi(K) = \det(sI - A - BKC).$$

The output pole placement problem has been studied by many authors. In 1975, Kimura [10] and Davison and Wang [5] proved independently that $m+p-1 \geqq n$ implies the generic pole assignability for generic systems. In 1977, by using the dominant morphism theorem, Hermann and Martin [8] proved that if $[A, X]=0$ and $CXB=0$ implies $X=0$, then the system has the generic pole-assignability by complex feedbacks. They also showed that $mp \geqq n$ implies the generic pole-assignability by complex feedbacks for generic systems. About a decade ago, inspired by a profound geometric construction due to Martin and Hermann [12] and by the questions raised by Willems and Hesslink [18], a geometric framework using Grassmannian variety was developed by Brockett and Byrnes [1], [3]. Since then, many new results have been proved using this framework. In 1981, Brockett and Byrnes [2] proved that when $mp = n$, if a system is nondegenerate, then the system has arbitrary pole-assignability by complex feedback and there are

$$d(m, p) = \frac{1! \cdots (p-1)!(mp)!}{m! \cdots (mp-1)!}$$

such feedback laws to assign each set of n poles. In 1983, by using the Ljusternik-Šnirel'mann category of real Grassmannians, Byrnes [4] proved that L-S$-$ $cat(\text{Grass}_R\ (p, m+p)) \geqq n+1$ implies generic pole assignability by real feedback that, when combined with the new results of L-S category of real Grassmannian, improved

* Received by the editors January 3, 1990; accepted for publication (in revised form) July 12, 1990.

† Department of Mathematics, Texas Tech University, Lubbock, Texas 79409.

Kimura's result. Recently, a new result that improved Kimura's result was proved by Rosenthal [14] using a very simple argument based on this framework.

One problem in feedback pole placement is to determine whether the pole placement map χ is dominant (see § 2). Note that χ is dominant over **C** if and only if it is dominant over **R**. When χ is dominant, it is almost onto over **C** by the dominant morphism theorem, and its image has nonempty interior over **R**.

Two necessary and sufficient conditions can be deduced from [8] and [18]. They are:

(a) χ is dominant if and only if there exists a K such that $[A+BKC, X]=0$ and $CXB=0$ implies that $X=0$.

(b) χ is dominant if and only if there exists a K such that $\{CB, C(A+BKC)B, \cdots, C(A+BKC)^{n-1}B\}$ are linearly independent.

In both conditions, the K must be determined before the conditions can be used.

In this paper, the dominant morphism theorem is applied to the central projection model of the pole placement map introduced by Byrnes [3]. The main result is that χ is dominant if and only if $mp \geqq n$, $\dim sp\{\text{minors of } G(s)\}=n$ and Grass $(p, m+p)$ is not contained in the Schubert variety $\sigma(E)$ under the embedding induced by the map $i: x \to C(T_x)$, where E is the center of χ. In the cases of 2×2 and 2×3 systems, if $mp=n$, the conditions become very simple: χ is dominant if and only if $\dim sp\{\text{minors of } G(s)\}=n$.

2. Preliminaries. Let X and Y be varieties, a morphism $\varphi: Y \to Y$ is called dominant if $\varphi(X)$ is Zariski-dense in Y. Let $k(X)$ and $k(Y)$ be the function fields of X and Y, then φ induces a homomorphism $\varphi: k(Y)\to k(X)$ by

$$\varphi^*(f(y))=f(y(x)).$$

It is easy to show that φ is dominant if and only if φ^* is one to one [7], [13], [15]. So if φ is dominant, $k(Y)$ can be considered as a subfield of $k(X)$.

PROPOSITION 2.1. *Let X and Y be affine varieties of the same dimensions over an algebraically closed field of characteristic* 0 *and $\varphi: X\to Y$ be a dominant morphism, then there is a Zariski-open set $Y_0 \subset Y$ such that $\#\varphi^{-1}(y)=[k(X): k(Y)]$ for all $y \in Y_0$.*

This is Proposition 3.17 of [13], except we are working over an arbitrary algebraically closed field of characteristic zero. The proof is the same.

The number $[k(X): k(Y)]$ is called the degree of φ.

PROPOSITION 2.2 (dominant morphism theorem). *Let $\varphi: X\to Y$ be a morphism of affine varieties over an algebraically closed field*; *then the following are equivalent*:

(i) *φ is dominant.*

(ii) *$\varphi(X)$ contains a nonempty Zariski open set of Y.*

(iii) *$d\varphi_x$ is onto for some smooth point x of X.*

(i)$\Rightarrow$(ii) is Theorem 6 of [15, § 5, Chap. I]. (i)$\Rightarrow$(iii) is Proposition 3.6 of [13]. (ii)$\Rightarrow$(i) is obvious. To prove (iii)$\Rightarrow$(i), suppose φ is not dominant. Let $Z=\overline{\varphi(X)}$; then $\dim Z < \dim Y$. So $\dim \operatorname{Im} d\varphi_x \leqq \dim Z < \dim Y$, $d\varphi_x$ cannot be onto for any x.

Let k be an algebraically closed field. A dominant morphism $\varphi: Y\to Y$ of affine varieties is called finite if $k[X]$ is integral over $k[Y]$. A finite morphism is epimorphic and it carries closed sets into closed sets of the same dimension [15]. A morphism $\varphi: X\to Y$ of quasi-projective varieties is called finite if every point $y\in Y$ has an affine neighborhood V such that the set $U=f^{-1}(V)$ is affine and $\varphi: U\to V$ is finite.

An example of finite morphism is the central projection. Let E be a d-dimensional subspace of a projective space $\mathbf{P}^n$, determined by $n-d$ linearly independent linear equations $L_1=L_2=\cdots=L_{n-d}=0$. The mapping $\pi(x)=(L_1(x), \cdots, L_{n-d}(x))$ is called a projection with the center at E; π is a morphism $\mathbf{P}^N - E \to \mathbf{P}^{n-d-1}$.

PROPOSITION 2.3. *If X is closed in $\mathbf{P}^n$ and $X \subset \mathbf{P}^n - E$, where E is a d-dimensional subspace, then the projection $\pi: X \to \mathbf{P}^{n-d-1}$ with the center at E determines a finite morphism $X \to \pi(X)$.*

The proof can be found in [15]. The geometrical meaning of the central projection is the following. As a model $\mathbf{P}^{n-d-1}$ we take any $(n-d-1)$-dimensional subspace $H \subset \mathbf{P}^m$ disjoint from E. Through any point $x \in \mathbf{P}^n - E$ and E there passes a unique $(d+1)$-dimensional subspace E_x. This subspace intersects H in a unique point, namely, $\pi(x)$. On the other hand, through any $y \in H$ and E there passes a unique $(d+1)$-dimensional subspace E_y. Then for all $x \in E_y \cap X - E$, $\pi(x) = y$. When $X \cap E = \emptyset$, π is finite, so $E_y \cap X$ has finite many points for all y. By Bezout's theorem (see [16, Prop. 3.26]) we have Proposition 2.4.

PROPOSITION 2.4. *If X is an $(n-d-1)$-dimensional subvariety of $\mathbf{P}^n$ and E is a d-dimensional subspace, $X \subset \mathbf{P}^n - E$, then the projection $\pi: X \to \mathbf{P}^{n-d-1}$with the center at E is onto and for any $y \in \mathbf{P}^{n-d-1}$,*

$$\# \pi^{-1}(y) = \deg X$$

counting multiplicity.

Now we introduce a projective variety, the Grassmannian Grass (m, n). Grass (m, n) is the variety of all *m-dimensional* subspaces in a *n-dimensional* vector space V^n. For each *m-space* $H \subset V^n$, if we choose a basis $\{\alpha_1, \cdots, \alpha_m\}$ and write each α_1 as column vector, then we have an $n \times m$ full rank matrix, $M_\alpha = [\alpha_1, \cdots, \alpha_m]$. For any $n \times m$ full rank matrix M, *col. sp* $M = H$ if and only if $M = M_\alpha T$ for some $T \in \mathrm{GL}(m)$. So we can say that Grass (m, n) is the variety of all $n \times m$ full rank matrices modulo $\mathrm{GL}(m)$. For any $n \times m$ matrix M of rank m, let $x_{i_1, \ldots, i_m}$ be the $m \times m$ minor of M by taking $i_1, \ldots, i_m$ rows of M. Since $x_{i_1, \ldots i_m}(MT) = x_{i_1, \cdots, i_m}(M) \det T$ for any $T \in \mathrm{GL}(m)$, we have an imbedding

$$p: \text{Grass}(m, n) \to \mathbf{P}^N,$$

where $N = \binom{n}{m} - 1$. This imbedding is called the Plücker imbedding. The imbedded Grassmannian is determined by a system of quadratic equations (see [6]).

3. Central projection and output feedback pole placement. Consider the closed-loop characteristic polynomial $\det(sI - A - BKC)$ of a linear system (A, B, C). It is well known that

$$\det(sI - A - BKC) = \det \begin{bmatrix} I & N(s) \\ K & D(s) \end{bmatrix},$$

where $G(s) = N(s) D^{-1}(s)$ is a right coprime factorization of the open-loop transfer function $G(s)$ [2], [3]. Since rank $\left[\begin{smallmatrix} I \\ K \end{smallmatrix}\right] = p$ and rank $\left[\begin{smallmatrix} N(s) \\ D(s) \end{smallmatrix}\right] = m$ for all s, $sp\left[\begin{smallmatrix} I \\ K \end{smallmatrix}\right]$ is a point in Grass $(p, m+p)$ and $sp\left[\begin{smallmatrix} N(s) \\ D(s) \end{smallmatrix}\right]$ is a curve in Grass $(m, m+p)$. We define the Plücker coordinates of Grass $(p, m+p)$ and Grass $(m, m+p)$ in the following way. Put all multiple index $(i_1, \cdots, i_p)$, $1 \leqq i_1 < \cdots < i_p \leqq m+p$ in order such that $(i_1, \cdots, i_p) < (j_1, \ldots, j_p)$ if there exists an s such that $i_t = j_t$ for all $t < s$ and $i_s < j_s$. For an ith index $(i_1, \cdots, i_p)$, let x_i be the $p \times p$ minor of $\left[\begin{smallmatrix} I \\ K \end{smallmatrix}\right]$ by using $i_1, \cdots, i_p$ rows and $p_i(s)$ be the $m \times m$ minor of $\left[\begin{smallmatrix} N(s) \\ D(s) \end{smallmatrix}\right]$ by eliminating $i_1, \cdots, i_p$ rows. Then

$$\det(sI - A - BKC) = \det \begin{bmatrix} I & N(s) \\ K & D(s) \end{bmatrix} \tag{1}$$

$$= \sum_{i=0}^{N} \theta(i) p_i(s) x_i, \tag{2}$$

where

$$N=\binom{m+p}{p}-1 \quad \text{and} \quad \theta(i)=(-1)^{p(p+1)/2+i_1+\cdots+i_p}$$

for the ith index $(i_1,\cdots,i_p)$. Note that $x_0=1$ and $p_0(s)=\det D(s)$. Let $\det(sI-A-BKC)=b_0+b_1s+\cdots+b_ns^n$ and $\theta(i)p_i(s)=a_{0i}+a_{1i}s+\cdots+a_{ni}s^n$. Then by equating coefficients of (1), we have

$$\begin{bmatrix} a_{00} & a_{01} & \cdots & a_{0N} \\ a_{10} & a_{11} & \cdots & a_{1N} \\ \vdots & \vdots & & \\ a_{n0} & a_{n1} & \cdots & a_{nN} \end{bmatrix}\begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_N \end{bmatrix}=\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} \tag{3}$$

or

$$Lx=b. \tag{4}$$

Let $E=\{x\in \mathbf{P}^N \mid Lx=0\}$; then (4) defines a central projection $\chi:\text{Grass}(p,m+p)-E\to \mathbf{P}^{r-1}$, where $r=\operatorname{rank} L$. So the pole placement map is the projection $\chi:\text{Grass}(p,m+p)-E\to\mathbf{P}^{r-1}$.

The geometrical meaning of χ is the following. Let $\mathbf{P}^n$ be the space of all polynomials of degree less than or equal to n with the equivalent relation $f\sim g$ if and only if $f=ag$ for some nonzero number α. Extending (4) to $\mathbf{P}^N$, we have $\tilde{\chi}:\mathbf{P}^N\to\mathbf{P}^n$ such that $\tilde{\chi}|_{\text{Grass}(p,m+p)}=\chi$, $\tilde{\chi}$ is a rational mapping and $\tilde{\chi}(\mathbf{P}^N)=\mathbf{P}^{r-1}\subseteq\mathbf{P}^n$. When we take any $(r-1)$-dimensional subspace $H\subset\mathbf{P}^N$ disjoint from E, $\tilde{\chi}|_H$ is one to one and $\tilde{\chi}(H)=\mathbf{P}^{r-1}$, so $H\cong\operatorname{Im}\tilde{\chi}$. We can identify the points in H with the points in $\operatorname{Im}\tilde{\chi}$. Through any point $x\in\text{Grass}(p,m+p)-E$ and E, there passes a unique $(N-r+1)$-dimensional subspace E_x. This subspace intersects H in a unique point, which is $\chi(x)$.

LEMMA 3.1.

$$\operatorname{rank} L=\dim sp\{\textit{minors of}\, G(s)\}+1,$$

where $sp\{minors\ of\, G(s)\}$ is a subspace spanned over $\mathbf{R}$ in $\mathbf{R}(s)$.

Proof. Since

$$\begin{bmatrix} N(s) \\ D(s)\end{bmatrix}=\begin{bmatrix} G(s) \\ I\end{bmatrix}D(s),$$

comparing the corresponding minors of each side, we have that

$$p_i(s)=m_i(s)\det D(s), \qquad i=1,2,\cdots,N,$$

where $\{m_i, i=1,2,\cdots,N\}$ are all minors of $G(s)$. So

$$\dim sp\{\text{minors of } G(s)\}=\dim sp\{p_i(s), i=1,2,\cdots,N\}.$$

Note that only $p_0(s)$ is a polynomial of degree n; all the other $p_i(s)$'s are polynomials of degree less than n. So

$$\begin{aligned}\dim sp\{p_i(s), i=1,2,\cdots,N\} &\\ &=\dim sp\{p_i(s), i=0,1,2,\cdots,N\}-1\\ &=\operatorname{rank} L-1.\end{aligned}$$

Recall the concept of nondegeneracy introduced by Brockett and Byrnes [2]: $G(s)$ is nondegenerate if and only if no hypersurface $\sigma(\omega)$ in Grass $(m, m+p)$ contains the curve

$$\begin{bmatrix} N(s) \\ D(s) \end{bmatrix},$$

where $\omega \in \text{Grass}\,(p, m+p)$ and

$$\sigma(\omega)=\{\nu \in \text{Grass}\,(m, m+p) \mid \dim\,(\omega \cap \nu) \geqq 1\}.$$

Now from the definition of E we can see that $G(s)$ is nondegenerate if and only if

$$E \cap \text{Grass}\,(p, m+p)=\phi.$$

When $G(s)$ is nondegenerate, the central projection χ is a finite morphism Grass $(p, m+p) \rightarrow \mathbf{P}^n$, since the degree of Grass $(p, m+p)$ is (see [9]):

$$d(m, p)=\frac{1!2!\cdots(p-1)!(mp)!}{m!(m+1)!\cdots(m+p-1)!}.$$

By Proposition 2.4, we have the Brockett-Byrnes theorem [3].

BROCKETT-BYRNES' THEOREM. *Let $G(s)$ be a nondegenerate $p \times m$ transfer function of McMillan degree $n=mp$. For all choices $(s_1, \cdots, s_n)$, we can find*

$$d(m, p)=\frac{1!2!\cdots(p-1)!(mp)!}{m!(m+1)!\cdots(m+p-1)!}$$

feedback laws over the complex field such that the closed-loop poles are exactly $(s_1, \cdots, s_n)$.

Actually, we have proved a more general result.

THEOREM 3.1. *If $G(s)$ is nondegenerate, then the image of χ : Grass $(p, m+p) \rightarrow \mathbf{P}^n$ is an irreducible projective variety of dimension mp. For each $p \in \operatorname{Im} \chi$, there are finite many feedback laws over the complex field such that the closed-loop characteristic polynomial is p.*

When $G(s)$ is degenerate, Grass $(p, m+p)-E$ is a quasi-projective variety and χ : Grass $(p, m+p) \rightarrow \mathbf{P}^n$ is a rational mapping. In this case, χ is almost onto over $\mathbf{C}$ if and only if χ is dominant. Here "almost onto" means that $\operatorname{Im} \chi$ contains a nonempty Zariski open set. We first prove a general result. Recall the concept of affine cone over a projective variety $X \subseteq \mathbf{P}^N$: For any nonempty algebraic set $X \subseteq \mathbf{P}^N$, let $\theta: \mathbf{A}^{N+1}-\{0\} \rightarrow \mathbf{P}^N$ be the mapping that sends the point with affine coordinates $(a_0, \cdots, a_N)$ to the point with homogeneous coordinates $(a_0, \cdots, a_N)$. The affine cone over X is defined as

$$C(X)=\theta^{-1}(X) \cup \{0\},$$

$$\dim C(X)=\dim X+1.$$

The tangent space $T_{x,X}$ of a projective variety X at x is defined as the projective closure of the tangent space of $X \cap U$ at x for any affine piece U of $\mathbf{P}^N$ that contains x.

PROPOSITION 3.1. *Let E be an $(N-n-1)$-dimension subspace of $\mathbf{P}^N$ over an algebraically closed field. For any projective variety $X \subset \mathbf{P}^N$, the central projection $\pi: X-E \rightarrow \mathbf{P}^n$ is almost onto if and only if $\dim X \geqq n$ and there exists a smooth point x of X such that*

$$\dim\,(T_{x,X} \cap E)=\dim X-n-1,$$

where the left-hand side equals -1 means that $T_{x,X} \cap E=\varnothing$.

Proof. Assume that E is defined by linear independent equations $L_0 = L_1 = \cdots = L_n = 0$. Define $\tilde{\pi}: C(\mathbf{P}^N) \to \mathbf{A}^{n+1}$ as

$$\tilde{\pi}(x) = (L_0(x), \cdots, L_n(x));$$

then $\theta(\tilde{\pi}(x)) = \pi(\theta(x))$ if $x \neq 0$ and $\tilde{\pi}(x) \neq 0$. $\pi: X - E \to \mathbf{P}^n$ is almost onto if and only if $\tilde{\pi}: C(X) \to \mathbf{A}^{n+1}$ is almost onto. By the dominant morphism theorem, $\tilde{\pi}: C(X) \to \mathbf{A}^{n+1}$ is almost onto if and only if $d\tilde{\pi}: T_{y,C(X)} \to \mathbf{A}^{n+1}$ is onto at some smooth point y of $C(X)$. Since $\tilde{\pi}$ is linear, $d\tilde{\pi}(T_{y,C(X)}) = \tilde{\pi}(T_{y,C(X)})$. Note that $y \neq 0$. Let $x = \theta(y)$; then

$$C(T_{x,X}) = T_{y,C(X)}.$$

Restrict $\tilde{\pi}$ on $C(T_{x,X})$,

$$\dim \operatorname{Im} \tilde{\pi} + \dim \ker \tilde{\pi} = \dim C(T_{x,X}) = \dim X + 1.$$

So $\tilde{\pi}: C(T_{x,X}) \to \mathbf{A}^{n+1}$ is onto if and only if

$$\dim \ker \tilde{\pi} = \dim X - n.$$

When $\dim X > n$, $\ker \tilde{\pi} \neq \{0\}$, $\theta(\ker \tilde{\pi} - \{0\}) = T_{x,X} \cap E$, so $\dim \ker \tilde{\pi} = n$ if and only if

$$\dim T_{x,X} \cap E = \dim X - n - 1.$$

When $\dim X = n$, $\ker \tilde{\pi} = \{0\}$ if and only if

$$T_{x,X} \cap E = \varnothing.$$

The theorem is proved.

Applying Proposition 3.1 to χ: Grass $(p, m+p) - E \to \mathbf{P}^n$, we have the following corollary.

COROLLARY 3.1. *If* $\dim E = \binom{m+p}{p} - n - 2 = N - n - 1$, *then* χ: Grass $(p, m+p) - E \to \mathbf{P}^n$ *is almost onto if and only if* $mp \geqq n$ *and there is an* $x \in$ Grass $(p, m+p)$ *such that*

$$\dim T_x \cap E = mp - n - 1 \quad \text{if } mp > n,$$

$$T_x \cap E = \varnothing \quad \text{if } mp = n.$$

For each x, T_x is an mp-dimensional subspace of $\mathbf{P}^N$, where $N = \binom{m+p}{p} - 1$, and $C(T_x)$ is an $(mp+1)$-dimensional linear subspace of $\mathbf{A}^{N+1}$. The map $i: x \to C(T_x)$ induces an embedding:

$$(5) \qquad i: \text{Grass}\,(p, m+p) \subset \text{Grass}\,(mp+1, N+1).$$

If $\dim E = \binom{m+p}{p} - n - 2 = N - n - 1$, $C(E)$ is an $(N-n)$-dimensional linear subspace of $\mathbf{A}^{N+1}$, so it is a point in Grass $(N-n, N+1)$.

Let

$$(6) \qquad \sigma(E) = \{T \in \text{Grass}\,(mp+1, N+1) \mid \dim (T \cap C(E)) \geqq mp - n + 1\}.$$

$\sigma(E)$ is a Schubert variety of codimension $mp - n + 1$ of Grass $(mp+1, N+1)$. It is the intersection of Grass $(mp+1, N+1)$ with a subspace of codimension $mp - n + 1$ of $\mathbf{P}^M$, where $M = \binom{N+1}{mp+1} - 1$ and Grass $(mp+1, N+1) \subset \mathbf{P}^M$ by Plücker embedding (see [11]), so we have Theorem 3.2.

THEOREM 3.2. *Let* $G(s)$ *be an* $m \times p$ *transfer function of McMillan degree* n, *then the output feedback pole placement map* χ: Grass $(p, m+p) - E \to \mathbf{P}^n$ *is dominant if and only if* $mp \geqq n$, $\dim$ *sp*$\{$*minors of* $G(s)\} = n$ *and* Grass $(p, m+p)$ *is not contained in the Schubert variety* $\sigma(E)$ *defined by* (6) *under the embedding induced by the mapping* $i: x \to C(T_x)$.

Proof. By Lemma 3.1, the condition of

$$\dim sp\{\text{minors of } G(s)\} = n$$

is equivalent to the condition of

$$\operatorname{rank} L = n+1$$

for the L defined in (4), which is equivalent to the condition of

$$\dim E = N - n - 1 = \binom{m+p}{p} - n - 2. \tag{7}$$

Note that (7) is a necessary condition. By Corollary 3.1, χ is almost onto if and only if $mp \geqq n$, $\dim sp\{\text{minors of } G(s)\} = n$ and there exists an $x \in \text{Grass}\,(m, m+p)$ such that

$$\dim C(T_x) \cap C(E) = mp - n.$$

Then by the definition of $\sigma(E)$, (6), the theorem is proved.

4. 2 × 2 systems of McMillan degree 4. We use our results to examine some systems. The simplest Grassmannian that is not a projective space is Grass (2, 4), which corresponds to the 2×2 systems of McMillan degree 4. The degree of Grass (2, 4) is 2, so if $G(s)$ is nondegenerate, χ is onto and 2 to 1 over **C**. For general cases, we have the following result.

THEOREM 4.1. *Let $G(s)$ be a 2×2 transfer function of McMillan degree 4. Then for almost all choices $(s_1, \cdots, s_4)$, we can find output feedback law over* **C** *such that the closed-loop poles are $(s_1, \cdots s_4)$ if and only if*

$$\dim sp\{\textit{minors of}\, G(s)\} = 4.$$

Proof. By Theorem 3.2, if we can prove

$$\text{Grass}\,(2,4) \not\subset \sigma(E)$$

for any E of dimension zero, then the theorem is proved. Grass (2, 4) is embedded in Grass (5, 6) and $\sigma(E)$ is a hyperplane section of Grass (5, 6). Since Grass $(5, 6) = \mathbf{P}^5$,

$$i: \text{Grass}\,(2,4) \subset \text{Grass}\,(5,6)$$

is the Plücker embedding, $\sigma(E)$ is a hyperplane so Grass $(2, 4) \not\subset \sigma(E)$ for any E of dimension zero.

Example 4.1. Let

$$G(s) = \frac{1}{s^4-1}\begin{bmatrix} s^3 & s \\ s & s^3 \end{bmatrix}.$$

The minors of $G(s)$ are $\{s^3/(s^4-1),\ s/(s^4-1),\ s^2/(s^4-1)\}$, so χ is not almost onto.

Example 4.2. Let

$$G(s) = \begin{bmatrix} 1/s & 1/s^2 \\ 0 & 1/s^3 \end{bmatrix}.$$

Then $G(s)$ is degenerate and the minors of $G(s)$ are $\{1/s,\ 1/s^2,\ 1/s^3,\ 1/s^4\}$,

$$\dim sp\{\text{minors of } G(s)\} = 4.$$

So we can place the poles of $G(s)$ to almost all choices $(s_1, \cdots, s_4)$ by complex output feedback. We can also place the poles to almost all self-conjugate $(s_1, \cdots, s_4)$ by real output feedback (see Theorem 4.2).

Now we consider the real field. Willems and Hesslink [18] proved that χ cannot be almost onto over $\mathbf{R}$ for generic system. Brockett and Byrnes [2] proved that if $G(s)$ is nondegenerate, χ is not almost onto over $\mathbf{R}$. So we consider the case when $G(s)$ is degenerate.

LEMMA 4.1. *If $G(s)$ is degenerate and χ is dominant, the degree of χ is one.*

Proof. Note that

$$\dim E = 0.$$

So if $G(s)$ is degenerate,

$$E \subset \text{Grass}\,(2, 4).$$

Let L be any line through E. If $L \not\subset \text{Grass}\,(2, 4)$, then L can intersect Grass $(2, 4)$ in at most one other point because

$$\deg \text{Grass}\,(2, 4) = 2.$$

Hence by Proposition 2.1,

$$\deg \chi = 1.$$

THEOREM 4.2. *Let $G(s)$ be a 2×2 transfer function of McMillan degree 4. Then for almost all self-conjugate $(s_1, \cdots, s_4)$ we can find a real output feedback law such that the closed-loop poles are $(s_1, \cdots, s_4)$ if and only if $G(s)$ is degenerate and*

$$\dim sp\{\textit{minors of}\, G(s)\} = 4.$$

Proof. χ is almost onto over $\mathbf{R}$ if χ is dominant and degree of χ is an odd number.

Brockett and Byrnes [2] proved that χ is onto over $\mathbf{R}$ if and only if $\det G(s) \equiv 0$ and

$$\dim sp\{g_1(s), g_2(s), g_3(s), g_4(s)\} = 4,$$

where $\{g_i\}$ are entries of $G(s)$. Note that $\det G(s) = 0$ is one of the cases of $G(s)$ being degenerate, so Brockett–Byrnes' result is included in Theorem 4.2.

5. 2×3 systems of McMillan degree 6. The next Grassmannian we will consider is the Grass $(2, 5)$. The degree of Grass $(2, 5)$ is 5, which is an odd number. So for a nondegenerate system, we can place poles arbitrarily over both the complex and the real fields.

For Grass $(2, 4)$, we have the property that Grass $(p, m+p) \not\subset \sigma(E)$ for any codimension $mp+1$ subspace E of $\mathbf{P}^N$, where $N = \binom{m+p}{p} - 1$ and Grass $(p, m+p)$ is embedded in Grass $(mp+1, N+1)$ (see § 3 and the proof of Theorem 4.1). For Grass $(2, 5)$, unfortunately, we do not have such a property. It is not difficult to find a subspace E of dimension two such that Grass $(2, 5) \subset \sigma(E)$, but, if we restrict such E to be the center of the projection χ for some system, it is not easy. After trying to construct such a system, we finally find that it is impossible. We have the same result as in Grass $(2, 4)$ as follows.

THEOREM 5.1. *Let $G(s)$ be a 2×3 transfer function of McMillan degree 6. Then χ is dominant if and only if*

$$\dim sp\{\textit{minors of}\, G(s)\} = 6.$$

We only give the ideal of the proof here. Please refer to [17] for details. The ideal is to find the forms of E such that Grass $(2, 5) \subset \sigma(E)$, then show that such E cannot be the center of χ for any transfer function $G(s)$.

Theorems 4.1 and 5.1 tempt us to conjecture that

$$\dim sp\{\text{minors of } G(s)\} = mp$$

might be the necessary and sufficient condition for χ being dominant for any $p\times m$ system of McMillan degree mp. However, this is not true. It is a necessary condition, but it is not sufficient; we must add the condition that Grass $(p, m+p)$ is not contained in the Schubert variety $\sigma(E)$ (see Theorem 3.2). We give an example of 3×3 transfer function $G(s)$ of McMillan degree 9 here.

Example 5.1. Let

$$G(s)=\frac{1}{s^4(s^3+1)(s-1)^2}\begin{bmatrix} s^6(s-1)^2 & -s^3(s-1)^2 & 0 \\ s^2(s-1)^2(s^2+1) & -s(s-1)^2(s^4+1) & 0 \\ s^4 & -s^5+s^4-s^2 & 0 \end{bmatrix}.$$

The dimension of $sp\{\text{minors of } G(s)\}$ equals 9, but the last column of $G(s)$ is zero. For any feedback law $u=Ky$, let

$$K=\begin{bmatrix} k_{11} & k_{12} & k_{13} \\ k_{21} & k_{22} & k_{23} \\ k_{31} & k_{32} & k_{33} \end{bmatrix}.$$

The values of k_{31}, k_{32}, k_{33} do not influence the closed-loop transfer function. So χ actually maps $\mathbf{A}^6$ into $\mathbf{A}^9$, it cannot be almost onto.

6. Conclusions. The output pole placement map χ has been studied by using a central projection model. It is proved that χ is dominant if and only if $mp\geqq n$, $\dim sp\{\text{minors of } G(s)\}=n$ and Grass $(p, m+p)\not\subset\sigma(E)$, where Grass $(p, m+p)$ is embedded in Grass $(mp+1, N+1)$ and $\sigma(E)$ is a Schubert variety of codimension $mp-n+1$ of Grass $(mp+1, N+1)$. In the cases of the 2×2 system of McMillan degree 4 and the 2×3 system of McMillan degree 6, it is proved that

$$\dim sp\{\text{minors of } G(s)\}=n$$

is a necessary and sufficient condition for χ being dominant, but this condition is not sufficient generally. It is also proved that for a 2×2 system of degree 4, χ is almost onto over the real field if and only if $G(s)$ is degenerate and

$$\dim sp\{\text{minors of } G(s)\}=4.$$

Acknowledgment. The author thanks C. I. Byrnes for his help and advice.

REFERENCES

[1] R. W. Brockett and C. I. Byrnes, *On the algebraic geometry of the output feedback pole placement map*, in Proc. 18th Conference on Decision and Control, Ft. Lauderdale, FL, 1979, pp. 754-757.

[2] ———, *Multivariable Nyquist criteria, root loci and pole placement: a geometric viewpoint*, IEEE Trans. Automat. Control, 26 (1981), pp. 271-284.

[3] C. I. Byrnes, *Algebraic and geometric aspects of the analysis of feedback systems*, in Geometric Methods in Linear Systems Theory, C. I. Byrnes and C. F. Martin, eds., D. Reidel, Dordrecht, the Netherlands, 1980, pp. 85-124.

[4] ———, *On the stability of multivariable systems and the Ljusternik-Šnivel'mann category of real Grassmannians*, Systems Control Lett., 3 (1983), pp. 255-262.

[5] E. J. Davison and S. H. Wang, *On pole assignment in linear multivariable systems using output feedback*, IEEE Trans. Automat. Control, 20 (1975), pp. 516-518.

[6] P. Griffiths and J. Adams, *Topics in Algebraic and Analytic Geometry*, Princeton University Press, Princeton, NJ, 1974.

[7] R. Hartshorne, *Algebraic Geometry*, Springer-Verlag, New York, 1977.

[8] R. Hermann and C. F. Martin, *Applications of algebraic geometry to systems theory, Part* 1, IEEE Trans. Automat. Control, 22 (1977), pp. 19-25.

[9] W. V. D. Hodge and D. Pedoe, *Methods of Algebraic Geometry, Vol.* II, Cambridge University Press, Cambridge, UK, 1952.

[10] H. KIMURA, *Pole assignment by gain output feedback*, IEEE Trans. Automat. Control, 20 (1975), pp. 509-516.

[11] S. L. KLEIMAN AND D. LAKSOV, *Schubert calculus*, Amer. Math. Monthly, 79 (1972), pp. 1061-1082.

[12] C. F. MARTIN AND R. HERMANN, *Applications of algebraic geometry to system theory: the McMillan degree and Kronecker indices as topological and holomorphic invariants*, SIAM J. Control Optim., 16 (1978), pp. 734-755.

[13] D. MUMFORD, *Algebraic Geometry* I: *Complex Projective Varieties*, Springer-Verlag, Berlin, New York, 1976.

[14] J. ROSENTHAL, *Tuning natural frequencies by output feedback*, in Computation and Control, K. Bowers and J. Lund, eds., Birkhaüser, Boston, 1989, pp. 277-282.

[15] I. R. SHAFAREVICH, *Basic Algebraic Geometry*, Springer-Verlag, Berlin, New York, 1977.

[16] W. VOGEL, *Lectures on results on Bezout's theorem*, Tata Institute of Fundamental Research, Bombay, India, 1984.

[17] X. WANG, *Additive inverse eigenvalue problems and pole placement of linear systems*, Ph.D. thesis, Department of Mathematics, Arizona State University, Tempe, AZ, 1989.

[18] J. C. WILLEMS AND W. H. HESSLINK, *Generic properties of the pole-placement problem*, in Proc. 7th IFAC Congress, 1978, Helsinki, Finland, pp. 1725-1728.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 936-953, July 1991

012

OPTIMAL CONTROL OF THE RUNNING MAX *

ARTHUR C. HEINRICHER† AND RICHARD H. STOCKBRIDGE‡

Abstract. A class of stochastic control problems where the payoff depends on the running maximum of a diffusion process is described. Such processes are appealing models for physical processes that evolve in a *continuous* and *increasing* manner. Dynamic programming conditions of optimality for these nonstandard problems are investigated and applied to particular examples.

Key words. controlled diffusion, running maximum, dynamic programming

AMS(MOS) subject classification. 93E20

1. Introduction.

This paper is devoted to the analysis of some nonstandard stochastic control problems which are derived from applications in which the state of the system is naturally monotone. They are closely related to standard controlled diffusion problems except that the payoff is allowed to depend on the *running maximum* of the diffusion. Our basic problem is of the following form.

The objective is to choose a control u_t to maximize

$$J(x,y;u) := E_{xy} \int_0^{\tau(B)} h(x_t, y_t, u_t)dt, \tag{1}$$

where the state is the pair (x_t, y_t) defined by the system

$$dx_t = f(x_t,u_t)dt + \sigma(x_t,u_t)dw_t, \quad x_0 = x, \tag{2}$$

$$y_t = \max\{x_s : 0 \leqq s \leqq t\} \vee y, \quad y_0 = y \geqq x. \tag{3}$$

We take $B \geqq y \geqq x$ and define $\tau(B)$ as the hitting time for x_t at level B. Here and throughout the paper, $w = (w_t,\ 0 \leqq t < \infty)$ denotes a standard, one-dimensional, Brownian motion.

To apply dynamic programming techniques to this problem, we consider the pair (x_t, y_t) as the *state.* (The process y_t alone is not a Markov process, but the pair (x_t, y_t) *is* a (strong) Markov process.) Pictured in the plane, the trajectories move on $y =$ const. lines while $x < y$, and there is motion in the y-direction only along the diagonal $x = y$.

The *value function V* reflects this picture of the state. It is defined in the half-plane $\{(x,y) : x \leqq y\}$. It satisfies a standard dynamic programming partial differential equation involving only derivatives with respect to x in the region $x < y$. The dependence on y_t in (1) becomes important only on the diagonal $x = y$ and adds a boundary condition to the dynamic programming equation. This picture also motivates the decomposition of the running max problem into a family of auxiliary control problems.

There is a large literature on optimal control for Markov processes, and diffusion processes in particular [13], [18], but it is impossible to model a *monotone* process via a (nondegenerate) diffusion because of the violent fluctuations of the Brownian

* Received by the editors September 5, 1989; accepted for publication (in revised form) July 23, 1990

† Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506-0027.

‡ Department of Statistics, University of Kentucky, Lexington, Kentucky 40506-0027.

motion. There is also a large literature working with monotone stochastic models in reliability theory and the statistics of lifetime distributions. Much of the work refers to the seminal paper of Esary, Marshall, and Proschan [12] on so-called *shock models.* These provide monotone models, but the paths increase in jumps and so are not suitable as models for phenomena that evolve in a *continuous* manner.

There is no literature on continuous, monotone stochastic processes simply because no nontrivial examples exist. In fact, Çinlar [7] has shown that the only *scalar* stochastic processes that are continuous, monotone, and strong Markov are *deterministic* functions of the initial state (up to a random killing). We can obtain a *vector* stochastic process that is continuous and strong Markov with one monotone component. By taking a diffusion and its running maximum, we obtain perhaps the simplest example of such a vector stochastic process.

One important application that requires a continuous and monotone stochastic process model is the optimal control of *wear*; tires, drill bits, and jet engines do not "unwear." Çinlar [5] has analyzed perhaps the most general class of stochastic processes suitable as models for wear (see [6], [8], [9]). In these *semi-Markov* processes, one component of a two-component process is monotone and assumed to model wear, while the second component is a Markov process referred to as the "excitation process." However, the generality of these processes detracts from their usefulness in certain control applications.

Baxter and Lee [3], [4] describe repair policies for a diffusion wear model. Conrad and McClamroch [10] use a controlled diffusion model for wear and describe applications to automated manufacturing. In both of these, the monotonicity of wear is surrendered to obtain a tractable control model.

The objective function (1) can be viewed with an interpretation similar to Conrad and McClamroch's [10] application in mind. Consider a control problem in which the state of a system is related to the level of wear in a machine and the control u can be interpreted as a *rate of working.* Profit is accrued while the machine is working, with an increase in the work rate providing an increase in the rate of income. The machine, however, will eventually deteriorate and fail (at level B in (1)), with increased wear decreasing the rate of income, and failure ending the process. (For such applications, the integrand $h(x, y, u)$ in (1) should be increasing as a function of u and decreasing as a function of x and y.) The controller is faced with competing objectives: he should work fast to maximize the rate of income, but working too fast will increase wear and will speed the failure of the machine.

1.1. Summary. We begin by describing sufficient conditions for optimality in terms of the Hamilton–Jacobi–Bellman (HJB) partial differential equation. In particular, we show that the appearance of the running maximum in the objective function is manifest only in the addition of an oblique derivative condition on the boundary of the halfplane $x \leqq y$.

We then describe how the running max problem can be solved by first solving a family of auxiliary problems. These are standard stochastic control problems and we give a simple formula for the running max value in terms of these auxiliary value functions.

One class of running max problems is particularly easy to solve. When the payoff depends *only* on the control and the running max, the optimal feedback control is constant for each of the auxiliary problems. This gives a simple formula for the auxiliary value function and hence for the running max value.

As an application, in §3 we solve a control problem in which the running profit depends in a linear way on the diffusion and its running maximum. The policy runs the gamut from a pure x-threshold policy when the payoff depends only on the diffusion process x to a pure y-threshold when the payoff depends only on the running max y. We investigate the regularity properties of this value function.

2. Dynamic programming conditions for optimality. We describe sufficient conditions for optimality in terms of the Hamilton–Jacobi–Bellman (HJB) partial differential equation. Our formulation is based on the standard verification theory as presented in Chapter VI of [13].

For a standard controlled diffusion problem, where the integrand in (1) does not depend on y_t and the state is given by (2), the (HJB) equation takes the form

$$\max_{u \in U}\left\{\frac{1}{2}\sigma(x,u)^2\, V''(x) + f(x,u)V'(x) + h(x,u)\right\} = 0 \tag{4}$$

for $x < B$ with the terminal condition

$$V(B) = 0. \tag{5}$$

If we can find a solution (in some appropriate sense) for this partial differential equation, then this solution will be the value function:

$$V(x) = \sup_{u \in \mathcal{A}} J(x;u).$$

We assume that controls take values in a compact set U and that the coefficients of the problem satisfy the following:

(A1) h is continuous and satisfies the polynomial growth condition

$$0 \leqq h(x,y,u) \leqq C\,(1 + |x|^p + |y|^p + |u|^p) \qquad (x \leqq y \leqq B,\ u \in U),$$

for suitable constants C and p;

(A2) f and σ are C^1 functions with

$$|f_x(x,u)| + |f_u(x,u)| \leqq K,\ |\sigma_x(x,u)| + |\sigma_u(x,u)| \leqq K\ (x \leqq B,\ u \in U),$$

and

$$|\sigma(x,u)| \leqq M\ (x \leqq B,\ u \in U),$$

for suitable constants K and M.

For the admissible controls, take the collection of *nonanticipative controls* as defined in Chapter VI of [13, p. 162]. Let $\mathcal{A}$ denote the collection of admissible controls.

The following example shows that the problem may be ill posed without a positive lower bound on $f(x,u)$. Consider a problem with control set $U = [0,1]$ and reward function

$$J(u) = E\int_0^{\tau(1)} \sqrt{u_t}\, dt,$$

where

$$dx_t = u_t dt + dw_t, \; x_0 = 0,$$

and $\tau(1)$ is the hitting time for level one. A maximizing sequence is obtained by taking constant controls $u_t^{(n)} \equiv 1/n$ for $n \geqq 1$. The state process is a drifted Brownian motion, with constant drift $1/n$, and so the expected time to reach level 1 is n. The expected reward is

$$J\left(u^{(n)}\right) = E\left[\int_0^{\tau(1)} \frac{1}{\sqrt{n}}\, dt\right] = \frac{1}{\sqrt{n}}\, E\left[\tau(1)\right] = \sqrt{n}.$$

There is no optimal policy; the controller can obtain arbitrarily large rewards by working arbitrarily slowly. We impose the following condition to preclude this behavior.

(A3) There is a constant α such that

$$0 < \alpha \leqq f(x,u) \; (x \leqq B, u \in U).$$

Our assumptions provide an exponential estimate on the probability that the diffusion exits the strip (A,B) at the lower boundary A. Let

$$\tau := \inf\left\{t \geqq 0 : x_t \notin (A,B)\right\}.$$

LEMMA 2.1. *Assume that $f(\cdot)$ and $\sigma(\cdot)$ satisfy assumptions* (A2) *and* (A3), *that $(x_t,\ 0 \leqq t)$ is a solution for* (2), *and that τ is defined as above. Then*

$$P_x\left(x_\tau = A\right) \leqq C \exp\left(kA\right), \tag{6}$$

where $k := 2\alpha/M^2 > 0$, and C depends on x, B, α, and M.

Proof. Define $\Phi(\cdot) : [A,B] \to [0,1]$ by

$$\Phi(x) = \frac{e^{-kx} - e^{-kB}}{e^{-kA} - e^{-kB}} \tag{7}$$

with k defined as above. Observe that

$$\Phi'(x) \leqq 0, \quad \Phi''(x) \geqq 0 \; (x \in (A,B)), \tag{8}$$

and that $\Phi(\cdot)$ satisfies the boundary value problem

$$\begin{gathered}\frac{1}{2}M^2\Phi''(x) + \alpha\Phi'(x) = 0, \; (x \in (A,B)), \\ \Phi(A) = 1, \; \Phi(B) = 0,\end{gathered} \tag{9}$$

By Itô's formula, taking expectations and using the fact that $\sigma(\cdot)$ and $\Phi'(\cdot)$ are bounded as well as the sign condition (8), we have

$$\begin{aligned} E_x\left[\Phi(x_\tau)\right] - \Phi(x) &= E_x \int_0^\tau \left[\frac{1}{2}\sigma^2(x_t,u_t)\Phi''(x_t) + f(x_t,u_t)\Phi'(x_t)\right] dt \\ &\leqq E_x \int_0^\tau \left[\frac{1}{2}M^2\Phi''(x_t) + \alpha\Phi'(x_t)\right] dt \\ &= 0. \end{aligned}$$

Therefore,

$$P_x(x_\tau = A) = E_x\left[\Phi(x_\tau)\right] \leqq \Phi(x) \leqq \left[\frac{e^{-kx} - e^{-kB}}{1 - e^{-kB}}\right] \exp(kA),$$

and the proof is compete. □

The following theorem is a direct extension of the standard sufficient condition for optimality.

THEOREM 2.2. *Let $V(x,y)$ be a solution of the dynamic programming equation*

$$\max_{u \in U}\left\{\frac{1}{2}\sigma(x,u)^2\, V_{xx}(x,y) + f(x,u)V_x(x,y) + h(x,y,u)\right\} = 0, \tag{10}$$

in the region $x < y < B$ satisfying the terminal condition

$$V(B,B) = 0, \tag{11}$$

as well as the boundary condition

$$V_y(y,y) = 0 \qquad (y \leqq B). \tag{12}$$

In addition, suppose $V(x,y)$ is continuous, twice continuously differentiable with respect to x, and satisfies a polynomial growth condition

$$|V(x,y)| \leqq C(1 + |x|^p + |y|^p) \qquad (x \leqq y,\ y \leqq B), \tag{13}$$

for appropriate constants C and p. Then:

(a) *$V(x,y) \geq J(x,y;u)$ for any admissible control u and any $x \leqq y$;*

(b) *If u^* is an admissible control that attains the maximum in* (10), *then u^* is optimal and $V(x,y) = J(x,y;u^*)$ is the value function.*

Proof. The proof is an application of the Itô formula. Assume that $V(x,y)$ satisfies (10)–(13). Let u be an admissible control and let (x_t, y_t) be the corresponding state process. Let T and N be positive constants and define

$$\tau_N := \inf\{t \geq 0 : x_t \leqq -N\}.$$

Then applying the Itô formula to $V(x_t, y_t)$, we have

$$\begin{aligned} V(x,y) \;&=\; -\int_0^{\tau(B)\wedge\tau_N\wedge T}\left[f(x_t,u_t)V_x(x_t,y_t) + \frac{1}{2}\sigma^2(x_t,u_t)V_{xx}(x_t,y_t)\right] dt \\ &\quad - \int_0^{\tau(B)\wedge\tau_N\wedge T} V_y(x_t,y_t)dy_t - \int_0^{\tau(B)\wedge\tau_N\wedge T} \sigma(x_t,u_t)V_x(x_t,y_t)dw_t \\ &\quad + \; V(x_{\tau(B)\wedge\tau_N\wedge T}, y_{\tau(B)\wedge\tau_N\wedge T}), \end{aligned}$$

where dy is the measure associated with the monotone increasing process y. Since the process $(y_t,\ 0 \leq t < \infty)$ increases only on the set $\{t : x_t = y_t\}$, the measure dy assigns mass only on this set. Hence the boundary condition (12) implies that the second integral is zero. Taking expectations and using (10) and the fact that the stochastic integral has zero expectation (the integrand is bounded), we obtain

$$\begin{aligned} V(x,y) &= -E_{xy}\int_0^{\tau(B)\wedge\tau_N\wedge T}\left[f(x_t,u_t)V_x(x_t,y_t)+\frac{1}{2}\sigma^2(x_t,u_t)V_{xx}(x_t,y_t)\right]dt \\ &\quad + E_{xy}V(x_{\tau(B)\wedge\tau_N\wedge T},y_{\tau(B)\wedge\tau_N\wedge T}) \\ &\geqq E_{xy}\int_0^{\tau(B)\wedge\tau_N\wedge T} h(x_t,y_t,u_t)dt + E_{xy}V(x_{\tau(B)\wedge\tau_N\wedge T},y_{\tau(B)\wedge\tau_N\wedge T}). \end{aligned}$$

The monotone convergence theorem implies that the integral term converges to the reward function $J(x,y;u)$ as T and $N\to\infty$. It follows from the continuity of V and the bounded convergence theorem that

$$\lim_{T\to\infty} E_{xy}V(x_{\tau(B)\wedge\tau_N\wedge T},y_{\tau(B)\wedge\tau_N\wedge T}) = E_{xy}V(x_{\tau(B)\wedge\tau_N},y_{\tau(B)\wedge\tau_N}).$$

To conclude, we must show that

$$(14)\qquad \lim_{N\to\infty} E_{xy}V(x_{\tau(B)\wedge\tau_N},y_{\tau(B)\wedge\tau_N}) = E_{xy}V(B,B)=0.$$

This will follow from the bound

$$(15)\qquad \sup_N E_{xy}\left[V^2(x_{\tau(B)\wedge\tau_N},y_{\tau(B)\wedge\tau_N})\right]<\infty,$$

because this implies uniform integrability.

The polynomial growth condition (13) provides the estimate

$$(16)\qquad E_{xy}\left[V^2(x_{\tau(B)\wedge\tau_N},y_{\tau(B)\wedge\tau_N})\right]\leqq C_1+C_2E_{xy}\left[x^{2p}_{\tau(B)\wedge\tau_N}\right],$$

where C_1 and C_2 depend on B, C, p, and the initial data. Because

$$E_{xy}\left[x^{2p}_{\tau(B)\wedge\tau_N}\right]\leqq B^{2p}+N^{2p}P_{xy}\left(\tau_N<\tau(B)\right),$$

Lemma 2.1, with $A=-N$, provides a bound for the right-hand side of (16) that is independent of N. Hence the family $V(x_{\tau(B)\wedge\tau_N},y_{\tau(B)\wedge\tau_N})$ is uniformly integrable. Equation (14) follows and we have

$$V(x,y)\geqq J(x,y;u)$$

for an arbitrary admissible control policy.

When u^* is an admissible control that attains the maximum in (10), the above argument holds with equality. □

Remark 2.3. If x_t is a Brownian motion, then the computations in [16, p. 95] can be modified to determine the transition densities for the Brownian motion and its running max (allowing arbitrary initial positions $x\leqq y$). We can then show that this process has infinitesimal generator

$$A\phi(x,y)=\frac{1}{2}\phi_{xx}(x,y)$$

with domain all functions defined in the halfplane $x \leqq y$, which are twice continuously differentiable with respect to x, continuously differentiable on the diagonal, and satisfy the boundary condition

$$\phi_y(x,y) = 0 \text{ when } x = y.$$

This derivation helps to explain, from another point of view, the origin of the boundary condition (12).

2.1. The auxiliary problem. The running max problem can be solved by first solving an *auxiliary problem.* In this problem y is a fixed parameter, the process starts at some $x \leqq y$, and the controller seeks to maximize the payoff obtained up to the first time the diffusion x_t reaches level y. This is a standard stochastic control problem for each $y \leqq B$.

In particular, let $x \leqq y$, define

$$\tau_1 = \tau_1(x,y;u) := \inf\{t \geq 0 : \ x_t = y\}, \tag{17}$$

and seek to maximize

$$J_1(x,y;u) := E_{xy} \int_0^{\tau_1} h(x_t, y, u_t)dt. \tag{18}$$

Let $W(x,y)$ denote the value function for this auxiliary problem:

$$W(x,y) := \sup_{u \in \mathcal{A}} J_1(x,y;u).$$

The HJB equation satisfied by $W(x,y)$ is of the standard form (4):

$$\max_{u \in U} \left\{ \frac{1}{2}\sigma(x,u)^2 \, W_{xx}(x,y) + f(x,u)W_x(x,y) + h(x,y,u) \right\} = 0 \tag{19}$$

on the halfline $x < y$ with the terminal condition

$$W(y,y) = 0. \tag{20}$$

There is a simple relationship between the value functions for the auxiliary problem and the running max problem. Using the *principle of optimality* (as presented in Lions [20, Thm. B]), we can show that the payoff starting at (x,y) consists of the auxiliary payoff $W(x,y)$ plus the payoff on the diagonal $V(y,y)$. Indeed, the principle of optimality says that for any admissible control and any stopping time θ,

$$V(x,y) = \sup_{u \in \mathcal{A}} E_{xy} \left[\int_0^{\theta \wedge \tau(B)} h(x_t, y_t, u_t)dt + V(x_{\theta\wedge\tau(B)}, y_{\theta\wedge\tau(B)}) \right].$$

Taking $\theta = \tau_1$, noting that $\tau_1 \leqq \tau(B)$ and $(x_{\tau_1}, y_{\tau_1}) = (y,y)$, this becomes

$$V(x,y) = \sup_{u \in \mathcal{A}} E_{xy}\left[J_1(x,y;u) + V(y,y)\right] = W(x,y) + V(y,y).$$

Thus we have the following proposition.

PROPOSITION 2.4. *The value functions for the running max problem and the auxiliary problem satisfy*

$$V(x,y) = W(x,y) + V(y,y) \qquad (x \leqq y, \ y \leqq B). \tag{21}$$

If we formally differentiate (21) with respect to y, evaluate on $x = y$, and use the boundary condition (12), we obtain

$$0 = V_y(y, y) = W_y(y, y) + \frac{d}{dy} V(y, y).$$

This indicates that the value for the running max problem can be represented entirely in terms of the value for the auxiliary problem.

THEOREM 2.5. *Let $W(x, y)$ be a solution of the dynamic programming equation* (19) *on the halfline $x < y$ satisfying the terminal condition* (20). *Suppose that $W(x, y)$ is twice continuously differentiable with respect to x and satisfies the polynomial growth condition*

$$|W(x, y)| \leqq C(1 + |x|^p + |y|^p) \qquad (x \leqq y),$$

for appropriate constants C and p. Then $W(x, y)$ is the value function for the auxiliary problem (18), *and if $u^*(x, y)$ is an admissible control that attains the maximum in* (19), *then $u^*(x, y)$ is an optimal control.*

In addition, assume that $W(x, y)$ is continuous with respect to (x, y) and differentiable along $x = y$. Then the running max value function is given by

$$V(x, y) = W(x, y) + \int_y^B W_y(z, z) dz \qquad (x \leqq y,\ y \leqq B). \tag{22}$$

The optimal control $u^*(x, y)$ for the auxiliary problem is also optimal for the running max problem (1).

Proof. The proof that $W(x, y)$ is the value function and that $u^*(x, y)$ is optimal for the auxiliary problem follows the standard verification theorem for a controlled diffusion (as in [13, Chap. VI]). (The proof is almost identical to that of Theorem 2.2, except that there is no integral with respect to dy_t to consider.)

Defining $V(x, y)$ as in (22), $V(x, y)$ inherits exactly the smoothness of $W(x, y)$; in particular, we have

$$V_x(x, y) = W_x(x, y),\ V_{xx}(x, y) = W_{xx}(x, y),$$

as well as

$$V(B, B) = W(B, B) = 0,$$

and

$$V_y(y, y) = W_y(y, y) - W_y(y, y) = 0.$$

Since $W(x, y)$ satisfies (19), $V(x, y)$ satisfies (10) as well as the terminal and boundary conditions (11) and (12). Theorem 2.2 identifies $V(x, y)$ as the value function and $u^*(x, y)$ as the optimal control policy for the running max problem. □

Remark 2.6. We have obtained *sufficient* conditions for optimality in terms of smooth, classical solutions to the (HJB) equation. The value function need not be a smooth classical solution, however. Recent work on existence and uniqueness in the class of generalized solutions known as *viscosity solutions* includes [11], [14], [15], [20], and [21]. In particular, Lions [20], [21] has developed dynamic programming conditions for standard control problems that do not require the smoothness hypothesized in Theorems 2.2 and 2.5. Barron [2] has applied these techniques and shown that the running max value function is the unique continuous viscosity solution for the boundary value problem (10)–(12).

2.2. The pure running max problem. When the payoff depends only on the control and the running max, and the drift and diffusion coefficients depend only on the control, the problem turns out to be remarkably simple. The reason for the simplification is that the optimal control is constant for fixed y and so the form of $W(x,y)$ is easy to derive.

Consider the following special running max problem. The objective is to choose a control u_t to maximize

$$J(x,y;u) := E_{xy}\int_0^{\tau(B)} h(y_t,u_t)dt, \tag{23}$$

where the state is given by

$$dx_t = f(u_t)dt + \sigma(u_t)dw_t,\ x_0 = x, \tag{24}$$

$$y_t = \max\{x_s : 0 \leqq s \leqq t\} \vee y. \tag{25}$$

Note that h depends only on y and u and the coefficients in the diffusion depend only on u.

When we consider the auxiliary objective function

$$J_1(x,y;u) = E_{xy}\int_0^{\tau_1} h(y,u_t)dt,$$

and restrict our attention to constant controls $u \in U$, the integrand is constant and thus

$$J_1(x,y;u) = h(y,u)E_{xy}\tau_1(x,y;u).$$

Furthermore, because the coefficients in (24) are constant, the expected value of the hitting time is known (see, for example, [17, §5, Chap. 7]):

$$E_{xy}\tau_1(x,y;u) = \frac{(y-x)}{f(u)}.$$

For a choice of $u \in U$, call it $u^*(y)$, which maximizes the ratio

$$u \mapsto \frac{h(y,u)}{f(u)}, \tag{26}$$

the payoff $J_1(x,y;u^*(y))$ satisfies the (HJB) equation (19) and the terminal condition (20). The standard verification theorem (as in [13, Chap. VI]) then identifies $J_1(x,y;u^*(y))$ as the value function $W(x,y)$ of the auxiliary problem and $u^*(y)$ as the optimal policy. The pair $(W(x,y),u^*(y))$ also satisfies the conditions of Theorem 2.5 and so $u^*(y)$ is the optimal policy for the pure running max problem and the value is given by (22). Summarizing, we have the following proposition.

PROPOSITION 2.7. *The value function for the auxiliary pure running max problem is*

$$W(x,y) = \frac{h(y,u^*(y))}{f(u^*(y))}(y-x), \tag{27}$$

where $u^*(y)$ maximizes the ratio (26). *The value function for the pure running max problem* (23) *is then*

$$V(x,y) = \frac{h(y,u^*(y))}{f(u^*(y))}(y-x) + \int_y^B \frac{h(z,u^*(z))}{f(u^*(z))}dz. \tag{28}$$

This explicit expression for $W(x,y)$ allows us to investigate its smoothness with respect to y. Clearly, if f, h, and u^* are twice continuously differentiable, $W(x,y)$ will be, also. The more interesting situation is when there is a jump in the optimal control $u^*(y)$.

Even if $u^*(y)$ is discontinuous, $W(x,y)$ will be continuous. To see this, assume that the optimal control $u^*(y)$ is identically α in a left-neighborhood of $\overline{y}$ and $u^*(y)$ is identically β in a right-neighborhood of $\overline{y}$. Then

$$\frac{h(\overline{y}^-,\beta)}{f(\beta)} \leqq \frac{h(\overline{y}^-,\alpha)}{f(\alpha)} = \frac{h(\overline{y}^+,\alpha)}{f(\alpha)} \leqq \frac{h(\overline{y}^+,\beta)}{f(\beta)} = \frac{h(\overline{y}^-,\beta)}{f(\beta)}.$$

(The inequalities follow from the optimality of α for $y < \overline{y}$ and of β for $y > \overline{y}$, and the equalities follow from the continuity of h.) As a result, equality holds throughout and continuity of h gives

$$\frac{h(\overline{y},\alpha)}{f(\alpha)} = \frac{h(\overline{y},\beta)}{f(\beta)} = \max_{u\in U}\frac{h(\overline{y},u)}{f(u)}. \tag{29}$$

So a switch occurs at an *indifference level*, a level $\overline{y}$ where the return obtained from control α is exactly the same as the return from control β. This is another way of saying that the value function (either the running max or the auxiliary value function) is continuous across the switching level.

When the optimal control does switch, there could be a jump in $W_y(x,y)$ at the switching level and we have a simple expression for the size of the jump. Let $\overline{y}$, α, and β be as above and assume that h is twice differentiable with respect to y. Then

$$\begin{aligned} \Delta W_y(x,\overline{y}) &:= W_y(x,\overline{y}^+) - W_y(x,\overline{y}^-) \\ &= h_y(\overline{y},\beta)E_{x\overline{y}}\,\tau_1(x,\overline{y};\beta) - h_y(\overline{y},\alpha)E_{x\overline{y}}\,\tau_1(x,\overline{y};\alpha) \qquad (30) \\ &= \left[\frac{h_y(\overline{y},\beta)}{f(\beta)} - \frac{h_y(\overline{y},\alpha)}{f(\alpha)}\right](\overline{y}-x). \end{aligned}$$

Note in particular the size of the jump decreases to zero at $x = \overline{y}$. In addition, the jump in W_{yy} is given by

$$\Delta W_{yy}(x,\overline{y}) = 2\left[\frac{h_y(\overline{y},\beta)}{f(\beta)} - \frac{h_y(\overline{y},\alpha)}{f(\alpha)}\right] + \left[\frac{h_{yy}(\overline{y},\beta)}{f(\beta)} - \frac{h_{yy}(\overline{y},\alpha)}{f(\alpha)}\right](\overline{y}-x). \tag{31}$$

The jumps in $W_y(x,y)$ and $W_{yy}(x,y)$ are inherited by $V_y(x,y)$ and $V_{yy}(x,y)$. For example,

$$\begin{aligned} \Delta V_y(x,\overline{y}) &:= V_y(x,\overline{y}^+) - V_y(x,\overline{y}^-) \\ &= \Delta W_y(x,\overline{y}) - \Delta W_y(\overline{y},\overline{y}) \\ &= \Delta W_y(x,\overline{y}). \end{aligned}$$

Remark 2.8. If we assume that

(a) $u \mapsto f(u)$ is positive, increasing, and concave on U; and

(b) $u \mapsto h(y,u)$ is positive, increasing, and convex for fixed y;

then for fixed x and y,

$$u \mapsto f(u)W_x(x,y) + h(y,u) \text{ is convex,}$$

and hence the maximum in the (HJB) equation is attained at the extreme values of the control set U. In other words, the optimal control is a *bang-bang* control.

3. The linear mixed problem. Consider now a control problem in which the control, the diffusion, and the running max enter the payoff in a linear way. Take

$$f(x,u) = u^\delta \text{ with } \delta \in (0,1], \text{ and } \sigma(x,u) \equiv 1,$$

so that the state is the pair (x_t, y_t) defined by the system

$$dx_t = (u_t)^\delta dt + dw_t,\ x_0 = x, \tag{32}$$

$$y_t = \max\{x_s : 0 \leqq s \leqq t\} \vee y, \tag{33}$$

with $x \leqq y$. The stopping time $\tau(B)$ is defined as before.

The admissible control processes $u = (u_t; 0 \leqq t < \infty)$ take values in

$$U := [\alpha, \beta] \text{ with } 0 < \alpha < \beta,$$

and the objective is to maximize the payoff

$$J(x,y;u) := E_{xy} \int_0^{\tau(B)} [cu_t - k_1 x_t - k_2 y_t] dt. \tag{34}$$

The parameters c, k_1, and k_2 are positive and we require that

$$c\beta - k_1 B - k_2 B \geqq 0. \tag{35}$$

This is sufficient to guarantee that the running profit can always be made positive. Alternatively, we can think of this last condition as *defining* the destruction level B as the maximum level at which it is possible to still make a profit.

We will construct the value function using the auxiliary problems described in §2 and show that the optimal control is a bang-bang control that switches between α and β at a switching line given by

$$k_1 x + k_2 y = \text{const.}$$

(see (41) below).

Remark 3.1. The restriction to the case $0 < \delta \leqq 1$ is made to give an optimal policy of the bang-bang form (recall Remark 2.8). When $\delta > 1$, the optimal policy is no longer simply α or β, but makes a smooth transition between these two values as x increases to y.

3.1. The auxiliary problem.

As described in §2.1, this running max problem can be solved by solving a family of simpler, standard control problems. Recall that the objective in the auxiliary problem is to maximize

$$J_1(x,y;u) := E_{xy} \int_0^{\tau_1} [cu_t - k_1 x_t - k_2 y] dt, \tag{36}$$

where $\tau_1 = \tau_1(x,y;u) := \inf\{t \geq 0 : \; x_t = y\}$, x_t is defined as in (32), and $y \geqq x$ is fixed.

The (HJB) equation (19) reduces to

$$\max_{\alpha \leqq u \leqq \beta} \left\{ \frac{1}{2} W_{xx}(x,y) + u^\delta W_x(x,y) + cu - k_1 x - k_2 y \right\} = 0 \tag{37}$$

on the halfline $x < y$ with the terminal condition

$$W(y,y) = 0. \tag{38}$$

We construct a solution $W(x,y)$ to (37) by first solving the equation with $u \equiv \alpha$ and then with $u \equiv \beta$, and patching these solutions together at a switching point $x = \overline{x}\,(y)$. (We are using the heuristic "principle of smooth fit." See Lehoczky and Shreve [19] for a discussion of this principle in singular and absolutely continuous control.)

The general solution for the (HJB) equation (37) with constant control (ignoring the maximization condition for a moment) is

$$\varphi(x,y;u) - \frac{c_1(u,y)}{2u^\delta} \exp\left(-2u^\delta x\right) + c_2(u,y), \tag{39}$$

where

$$\varphi(x,y;u) := \frac{k_1}{2u^\delta} x^2 - \left(\frac{cu}{u^\delta} + \frac{k_1}{2u^{2\delta}} \right) x + \frac{k_2}{u^\delta} xy. \tag{40}$$

We have one solution for $u = \alpha$ and one for $u = \beta$. The five parameters: $c_1(\alpha, y)$, $c_1(\beta, y)$, $c_2(\alpha, y)$, $c_2(\beta, y)$, and the switching point $\overline{x}\,(y)$, are at our disposal. These five parameters are fixed by the terminal condition (38), continuity of the function and its first two derivatives, and a growth condition for x large and negative.

Standard techniques provide the following estimate for the growth of $W(x,y)$.

LEMMA 3.2. *There is a positive constant $C = C(\alpha, \beta, \delta, c, k_1, k_2)$ such that*

$$0 \leqq \sup_{u \in \mathcal{A}} J_1(x,y;u) \leqq C\left(1 + |x|^2 + |y|^2\right) \; (x \leqq y).$$

In particular, the polynomial growth condition (13) *is satisfied with $p = 2$.*

This lemma rules out exponential growth for x large and negative. The fact that $\varphi(x,y;\alpha) > \varphi(x,y;\beta)$ for x large and negative indicates that the optimal control must be $u = \alpha$ for such x, and so $c_1(\alpha, y) \equiv 0$. Hence

$$W_x(x,y) = \frac{k_1}{\alpha^\delta} x - \left(\frac{c\alpha}{\alpha^\delta} + \frac{k_1}{2\alpha^{2\delta}} \right) + \frac{k_2}{\alpha^\delta} y \; (x < \overline{x}\,(y)).$$

For $x > \overline{x}\,(y)$, we use $u = \beta$, so

$$W_x(x,y) = \frac{k_1}{\beta^\delta} x - \left(\frac{c\beta}{\beta^\delta} + \frac{k_1}{2\beta^{2\delta}} \right) + \frac{k_2}{\beta^\delta} y + c_1(\beta, y) \exp\left(-2\beta^\delta x\right) \; (\overline{x}\,(y) < x < B).$$

If we require that $W_x(x,y)$ be continuous across $x = \overline{x}\,(y)$, we have one restriction:

$$\begin{aligned}&\tfrac{k_1}{\alpha^\delta}\overline{x}\,(y) - \left(\tfrac{c\alpha}{\alpha^\delta} + \tfrac{k_1}{2\alpha^{2\delta}}\right) + \tfrac{k_2}{\alpha^\delta}y \\ &\quad = \tfrac{k_1}{\beta^\delta}\overline{x}\,(y) - \left(\tfrac{c\beta}{\beta^\delta} + \tfrac{k_1}{2\beta^{2\delta}}\right) + \tfrac{k_2}{\beta^\delta}y + c_1(\beta,y)\exp\left(-2\beta^\delta\overline{x}\,(y)\right).\end{aligned}$$

Continuity of $W_{xx}(x,y)$ across $x = \overline{x}\,(y)$ provides a second restriction:

$$\frac{k_1}{\beta^\delta} - 2\beta^\delta c_1(\beta,y)\exp\left(-2\beta^\delta\overline{x}\,(y)\right) = \frac{k_1}{\alpha^\delta}.$$

Combining these we obtain the switching line

$$\overline{x}\,(y) = \frac{1}{2\alpha^\delta} + \frac{c}{k_1}\left(\frac{\alpha\beta^\delta - \alpha^\delta\beta}{\beta^\delta - \alpha^\delta}\right) - \frac{k_2}{k_1}y, \tag{41}$$

as well as

$$c_1(\beta,y) = -\frac{k_1}{2\beta^\delta}\left(\frac{\beta^\delta - \alpha^\delta}{\alpha^\delta\beta^\delta}\right)e^{2\beta^\delta\overline{x}\,(y)}. \tag{42}$$

Note that there is a switch (i.e., $\overline{x}\,(y) < y$) if and only if

$$y > \hat{y} := \frac{1}{2\alpha^\delta}\left(\frac{k_1}{k_1+k_2}\right) + \frac{c}{k_1+k_2}\left(\frac{\alpha\beta^\delta - \alpha^\delta\beta}{\beta^\delta - \alpha^\delta}\right). \tag{43}$$

The constants $c_2(\beta,y)$ and $c_2(\alpha,y)$ are determined by requiring $W(y,y) = 0$ and continuity across $\overline{x}\,(y)$ for $y > \hat{y}$. In particular, for $y \leqq \hat{y}$ there is no switch and so

$$c_2(\alpha,y) = -\varphi(y,y;\alpha)\ (y \leqq \hat{y}). \tag{44}$$

For $y > \hat{y}$, the control is β for $\overline{x}\,(y) < x \leqq y$ and $W(y,y) = 0$ provides

$$c_2(\beta,y) = -\frac{k_1}{4\beta^{2\delta}}\left(\frac{1}{\alpha^\delta} - \frac{1}{\beta^\delta}\right)\exp\left(-2\beta^\delta(y - \overline{x}\,(y))\right) - \varphi(y,y;\beta). \tag{45}$$

Continuity at $x = \overline{x}\,(y)$ requires

$$c_2(\alpha,y) = -\varphi(\overline{x}\,(y),y;\alpha) + \varphi(\overline{x}\,(y),y;\beta) + \frac{k_1}{4\beta^{2\delta}}\left(\frac{1}{\alpha^\delta} - \frac{1}{\beta^\delta}\right) + c_2(\beta,y). \tag{46}$$

This construction provides a solution to (37), which is twice continuously differentiable (with respect to x) and also satisfies the growth condition of Lemma 3.2. The standard verification theorem for controlled diffusions provides the following proposition.

PROPOSITION 3.3. *The optimal control for the auxiliary problem* (36) *is*

$$u^*(x,y) = \begin{cases} \alpha, & x < y,\ y \leqq \hat{y}, \\ \alpha, & x < \overline{x}\,(y),\ y > \hat{y}, \\ \beta, & \overline{x}\,(y) < x \leqq y,\ y > \hat{y}, \end{cases} \tag{47}$$

where $\overline{x}\,(y)$ is defined in (41) *and $\hat{y}$ is defined in* (43). *The value function is given by*

$$
(48)W(x,y)=\begin{cases}\varphi(x,y;\alpha)-\varphi(y,y;\alpha),\ x\leqq y,\ y\leqq\hat{y},\\[2ex] \varphi(x,y;\alpha)+c_2(\alpha,y),\ x\leqq\overline{x}\,(y),\ y>\hat{y},\\[2ex] \varphi(x,y;\beta)-\dfrac{c_1(\beta,y)}{2\beta^\delta}\,e^{(-2\beta^\delta x)}+c_2(\beta,y),\ \overline{x}\,(y)<x\leqq y,\ y>\hat{y},\end{cases}
$$

where $\varphi(x,y;u)$ is defined in (40) *and the parameters $c_2(\alpha,y)$, $c_1(\beta,y)$, and $c_2(\beta,y)$ are defined in* (46), (42), *and* (45), *respectively.*

By construction, $W(x,y)$ is twice continuously differentiable with respect to x. As a function of y, $W(x,y)$ is equally well behaved *as long as* $k_1>0$. In particular, for fixed x, the control switches from α to β at

$$
(49)\qquad \overline{y}\,(x):=\frac{1}{2\alpha^\delta}\frac{k_1}{k_2}+\frac{c}{k_2}\left(\frac{\alpha\beta^\delta-\alpha^\delta\beta}{\beta^\delta-\alpha^\delta}\right)-\frac{k_1}{k_2}x.
$$

We verify that $W_y(x,y)$ is continuous across $y=\overline{y}\,(x)$; the computations to verify the continuity of $W(x,y)$ and $W_{yy}(x,y)$ are similar. Looking at $W_y(x,y)$ above and below $y=\overline{y}\,(x)$ (letting "$'$" denote differentiation with respect to y), we obtain

$$
\begin{aligned}
\Delta W_y(x,\overline{y}\,(x)) &:= W_y(x,\overline{y}\,(x)^+)-W_y(x,\overline{y}\,(x)^-)\\
&= \varphi_y(x,\overline{y}\,(x);\beta)-\left(c_1'(\beta,\overline{y}\,(x))\frac{1}{2\beta^\delta}\right)e^{-2\beta^\delta x}\\
&\quad+c_2'(\beta,\overline{y}\,(x))-\varphi_y(x,\overline{y}\,(x);\alpha)-c_2'(\alpha,\overline{y}\,(x))\\
&= \left[\frac{k_1}{2\beta^\delta}\left(\frac{\beta^\delta-\alpha^\delta}{\beta^\delta\alpha^\delta}\right)+\varphi_x(x,\overline{y}\,(x);\alpha)-\varphi_x(x,\overline{y}\,(x);\beta)\right]\overline{x}\,'(y)\\
&= \left[\frac{k_1}{2\beta^\delta}\left(\frac{\beta^\delta-\alpha^\delta}{\beta^\delta\alpha^\delta}\right)-\frac{k_1}{2\beta^\delta}\left(\frac{\beta^\delta-\alpha^\delta}{\beta^\delta\alpha^\delta}\right)\right]\overline{x}\,'(y)\\
&= 0.
\end{aligned}
$$

We have used the definitions of $c_1(\beta,y)$, $c_2(\alpha,y)$, $c_2(\beta,y)$ given above as well as $\overline{x}\,(\overline{y}\,(x))=x$ and

$$
\left[c_1(\beta,\overline{y}\,(x))e^{-2\beta^\delta x}+\varphi_x(x,\overline{y}\,(x);\beta)-\varphi_x(x,\overline{y}\,(x);\alpha)\right]=W_x(x,\overline{y}\,(x)^+)-W_x(x,\overline{y}\,(x)^-)=0.
$$

Note that

$$
\overline{x}\,'(y)=-\frac{k_2}{k_1}
$$

becomes infinite as $k_1\to 0$. This corresponds with the appearance of the jump in $W_y(x,y)$ at $y=\overline{y}$ (recall (30)) for the pure running max problem ($k_1=0$).

Since $W(x,y)$ is continuous in the region $\{(x,y)\,:\,x\leqq y\}$, twice continuously differentiable with respect to x, continuously differentiable on the diagonal $x=y$, and has polynomial growth, Theorem 2.5 can be applied to construct the value function for the running max problem. The optimal policy for the running max problem is the same as the policy described in Proposition 3.3.

PROPOSITION 3.4. *The optimal feedback control for the linear mixed problem* (34) *is given in* (47). *The value function is given by*

$$V(x,y) = W(x,y) + \int_y^B W_y(z,z)dz, \tag{50}$$

with $W(x,y)$ *defined in Proposition* 3.3, (48).

As long as $k_1 > 0$, $V(x,y)$ is also twice continuously differentiable in both x and y. In fact, $V(x,y)$ inherits its smoothness from $W(x,y)$ since

$$V_y(x,y) = W_y(x,y) - W_y(y,y),$$

and

$$V_{yy}(x,y) = W_{yy}(x,y) - W_{xy}(y,y) - W_{yy}(y,y).$$

3.2. Two special cases. We now discuss two limiting cases for the mixed problem.

3.2.1. The standard problem: $\mathbf{k_2 = 0}$. If the running maximum y_t does not enter the objective function, then the control problem is reduced to a standard optimal stochastic control problem.

The switching level defined in (41) becomes constant:

$$\overline{x} = \frac{1}{2\alpha^\delta} + \frac{c}{k_1}\left(\frac{\alpha\beta^\delta - \alpha^\delta\beta}{\beta^\delta - \alpha^\delta}\right) = \frac{1}{2\alpha^\delta} + \frac{c\alpha}{k_1}\left[1 - \left(\frac{(\beta/\alpha) - 1}{(\beta/\alpha)^\delta - 1}\right)\right]. \tag{51}$$

Note that the term in parentheses is negative when $0 < \alpha < \beta$ and $0 < \delta < 1$.

The switching level $\overline{x}$ defined in (51) is decreased if the ratio c/k_1 is increased. That is, if the coefficient of u is increased in the running cost, then it is optimal to use the maximum work rate longer. Similarly, if c/k_1 is small, so the coefficient of x in the running cost dominates, then $\overline{x}$ approaches the maximum switching level $1/2\alpha^\delta$. Finally, in the case where $\delta = 1$, the switching level becomes independent of the cost parameters c and k_1, since $\overline{x} = 1/(2\alpha)$.

3.2.2. The pure running max version: $\mathbf{k_1 = 0}$. If the diffusion x_t does not enter the objective function, then this is a pure running max problem, so we know that the optimal policy will be a y-threshold policy. In particular, the switching point $\overline{x}(y)$ defined in (41) becomes $-\infty$ and $\overline{y}(x)$ defined in (49) becomes constant:

$$\overline{y} = \frac{c}{k_2}\left(\frac{\alpha\beta^\delta - \alpha^\delta\beta}{\beta^\delta - \alpha^\delta}\right) = \frac{c\alpha}{k_2}\left[1 - \left(\frac{(\beta/\alpha) - 1}{(\beta/\alpha)^\delta - 1}\right)\right]. \tag{52}$$

The optimal policy is

$$u^*(x,y) = \begin{cases} \alpha, & y \leqq \overline{y}; \\ \beta, & \overline{y} < y \leqq B. \end{cases} \tag{53}$$

Note that $\overline{y}$ is negative for $0 < \delta < 1$, increasing to zero when $\delta = 1$, hence it is never optimal to use the slow work rate if the initial state y is nonnegative.

Remark 3.5. One of the original motivations for this research was to determine what happens when the objective for a controlled diffusion problem depends on the running maximum instead of the diffusion. For the linear example, we have shown that the policy changes from an x-threshold to a y-threshold policy and the switching level for the running max problem is *lower* than the switching level for the standard problem. That is, let $\overline{x}$ denote the optimal switching level for the payoff

$$J(x;u) := E_x \int_0^{\tau(B)} [cu_t - kx_t]dt$$

and let $\overline{y}$ denote the switching level for

$$J_{\max}(x,y;u) := E_{xy} \int_0^{\tau(B)} [cu_t - ky_t]dt.$$

Then

$$\overline{y} = \overline{x} - \frac{1}{2\alpha^\delta} < \overline{x}.$$

As noted in §2.2, the value function for a pure running max problem does not retain the smoothness (with respect to y) demonstrated after Proposition 3.3. The value function is given by

$$V(x,y) = W(x,y) + \int_y^B W_y(z,z)dz,$$

where, recalling (27),

$$W(x,y) = \begin{cases} (c\alpha - k_2 y)\dfrac{(y-x)}{\alpha^\delta}, & y \leqq \overline{y}, \ x \leqq y, \\ (c\beta - k_2 y)\dfrac{(y-x)}{\beta^\delta}, & \overline{y} < y \leqq B, \ x \leqq y. \end{cases} \tag{54}$$

Note that $W(x,y)$ satisfies the the hypotheses of Theorem 2.5.

The value function $V(x,y)$ is not continuously differentiable with respect to y; there is a "corner" at $\overline{y}$. Recalling (30), $V_y(x,y)$ has a positive jump at $\overline{y}$ of magnitude

$$\begin{aligned} \Delta V_y(x,\overline{y}) &= \Delta W_y(x,\overline{y}) \\ &= k_2 \left[E_{x\overline{y}}\, \tau_1(x,\overline{y}\,;\beta) - E_{x\overline{y}}\, \tau_1(x,\overline{y}\,;\alpha)\right] \\ &= k_2 \left(\frac{\beta^\delta - \alpha^\delta}{\alpha^\delta \beta^\delta}\right)(\overline{y} - x). \end{aligned}$$

The jump in $V_y(x,y)$ is just k_2 times the jump in the expected hitting time when the control switches.

Finally, there is a jump in $V_{yy}(x,y)$ at $y = \overline{y}$ of magnitude

$$\Delta V_{yy}(x,\overline{y}) = 2k_2 \left(\frac{\beta^\delta - \alpha^\delta}{\beta^\delta \alpha^\delta}\right).$$

4. Concluding remarks. The running max provides a continuous, monotone stochastic process model with the advantages of all the machinery of the Itô calculus for applications. The theory of controlled diffusions extends to handle problems in controlling the running max. It remains to be seen in applications whether a running max model will provide results fundamentally different from results for standard controlled diffusions.

One important application for running max problems is in optimal control and replacement problems for systems subject to deterioration and wear. If we take the linear mixed problems considered in §3 and consider it as such a problem, we obtain a quite counterintuitive policy: work cautiously ($u = \alpha$) when the machine is new and work the machine very hard ($u = \beta$) as the machine ages and wears. This behavior is an artifact of the form of the objective function and not the running max model for the state. In particular, there is a large penalty for working with a worn machine and there is no penalty for failure while working, hence there is no reward for working conservatively with a worn machine. There is also no option to quit before failure and there is no reward for doing so. Thus, the linear payoff does not fit the form usually studied in optimal repair and replacement problems (see [1], [10], [12], and [22] for example). In future work, we will consider running max control problems where the objective function is closer to the types of objectives usually considered in these applications.

There are interesting and difficult estimation and identification problems introduced in applications of the running max model. Recall that it was not possible to simply *replace* the diffusion with the running max; the running max alone is not a Markov process. It was necessary to assume that the controller could observe *both* the diffusion $(x_t; 0 \leqq t < \infty)$ *and* the running maximum $(y_t; 0 \leqq t < \infty)$. How is the problem changed, indeed is it solvable in any sense, if only the values of the running maximum are available to the controller? As shown in §2.2, if the diffusion does not enter the payoff function directly, then the optimal policy depends only on the observed values of y_t. However, if both the diffusion and its running maximum enter the payoff, the optimal policy depends on the observed value of x_t, and then we are faced with a *partially observed* stochastic control problem, and the problems become much more difficult.

Acknowledgments. We thank Professors Kurt Helmes and Raymond Rishel for helpful comments and criticisms during the development of this research. We also thank Professors Steve Shreve and Tom Kurtz for many helpful discussions.

REFERENCES

[1] R. S. Anderson, *Replacement with nonconstant operating cost*, SIAM J. Control Optim., 26 (1988), pp. 1076–1098.

[2] E. M. Barron, *The Bellman equation for control of the running max of a diffusion and applications to look-back options*, preprint.

[3] L. A. Baxter and E. Y. Lee, *A diffusion model for a system subject to continuous wear*, Probab. Engrg. Inform. Sci., 1 (1987), pp. 405–416.

[4] ———, *Optimal control of a model for a system subject to continuous wear*, Probab. Engrg. Inform. Sci., 2 (1988), pp. 321–328.

[5] E. Çinlar, *Markov additive processes,* I & II, Z. Wahrsch. Verw. Gebiete., 24 (1972), pp. 85–121.

[6] ———, *Shock and wear models and Markov additive processes*, in Theory and Applications of Reliability, Vol. 1, Academic Press, New York, 1977.

[7] ———, *On increasing continuous processes*, Stochastic Process. Appl., 9 (1979), pp. 147–154.

[8] E. ÇINLAR, *Markov and Semi-Markov models for deterioration*, in Reliability Theory and Models, Academic Press, New York, 1984.

[9] E. ÇINLAR AND S. ÖZEKICI, *Reliability of complex devices in random environments*, Probab. Engrg. Inform. Sci., 1 (1987) pp. 97–115.

[10] C. CONRAD AND N. H. MCCLAMROCH, *The drilling problem: A stochastic modelling and control example in manufacturing*, IEEE Trans. Automat. Control, 32 (1987), pp. 947–958.

[11] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton–Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1–42.

[12] J. D. ESARY, A. W. MARSHALL, AND F. PROSCHAN, *Shock models and wear processes*, Ann. Probab., 1 (1973), pp. 627–649.

[13] W. H. FLEMING AND R. W. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer–Verlag, New York, Heidelberg, Berlin, 1975.

[14] H. ISHII, *On uniqueness and existence of viscosity solutions of nonlinear second order elliptic* PDE*s*, preprint.

[15] R. JENSEN, *The maximum principle for viscosity solutions of fully nonlinear second order partial differential equations*, Arch. Rational Mech. Anal., 101 (1988), pp. 1–27.

[16] I. KARATZAS AND S. E. SHREVE, *Brownian Motion and Stochastic Calculus*, Springer-Verlag, New York, Heidelberg, Berlin, 1987.

[17] S. KARLIN AND H. M. TAYLOR, *A First Course in Stochastic Processes*, Academic Press, New York, San Francisco, London 1975.

[18] N. V. KRYLOV, *Controlled Diffusion Processes*, Springer-Verlag, New York, Heidelberg, Berlin, 1980.

[19] J. P. LEHOCZKY AND S. E. SHREVE, *Absolutely continuous and singular stochastic control*, Stochastics, 17 (1986), pp. 91–109.

[20] P. L. LIONS, *Optimal control of diffusions and Hamilton–Jacobi equations. Part* I: *the dynamic programming principle and applications*, Comm. Partial Differential Equations, 8 (1983), pp. 1101–1174.

[21] ———, *Optimal control of diffusions and Hamilton–Jacobi equations. Part* II: *viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1229–1276.

[22] Y. S. SHERIF AND M. L. SMITH, *Optimal maintenance models for systems subject to failure—a review*, Naval Res. Logist. Quart., 28 (1981), pp. 47–74.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 954-967, July 1991

013

FINITE TIME OBSERVER DESIGN BY PROBABILISTIC-VARIATIONAL METHODS*

MATTHEW R. JAMES†

Abstract. A notion of finite time observer for partially observed deterministic control systems is introduced; it is shown how such observers are obtained by probabilistic-variational methods. Observability plays an essential role. The procedure is carried out for finite-state discrete-time systems and continuous-time nonlinear systems, and the observers obtained are respectively finite and infinite dimensional. A simple algorithm implementing the observer for finite-state systems is described. An observability grammian for nonlinear systems is introduced and is used to study the time evolution of sets of indistinguishable points, as well as local properties of the function satisfying the observer equation. Finally, the results are specialized to bilinear systems.

Key words. nonlinear control systems, observers, filters, observability

AMS(MOS) subject classifications. 93B07, 93E11

1. Introduction.

In this paper we introduce a notion of *finite time observer* for partially observed deterministic control systems and show how such observers are obtained by probabilistic-variational methods. A finite time observer is a dynamical system that uses observations from a control system to compute an estimate $\hat{x}(t)$ of the state $x(t)$ of the control system; and after a finite time has elapsed, the estimate is exact: for any *universal* input defined on $[0,T]$, $\hat{x}(t) = x(t)$, $t \geq T$. Most existing observer designs are *asymptotic*: $\text{dist}(\hat{x}(t), x(t)) \to 0$ as $t \to \infty$.

In [3], Baras and Krishnaprasad proposed that observers for nonlinear control systems might be obtained as asymptotic limits of recursive filters. Their idea was to add noise, scaled by a small parameter $\epsilon > 0$, to the equations defining the system, construct the corresponding family of filters parameterized by $\epsilon > 0$, and then pass to the limit as $\epsilon \to 0$. The limiting filter that results is a candidate observer for the original deterministic system. Finally, we must determine whether the candidate is, in fact, an observer. Baras [1] conjectured that, in general, an "infinite-dimensional" observer would result, under appropriate hypotheses (such as observability). The rationale behind this idea was to achieve an observer design by exploiting both the additional structure obtained by randomizing the problem, and the power of asymptotic methods.

To date, work on this approach has been primarily concerned with studying the asymptotic filtering problem (James and Baras [17], James [15]) and using the limiting filter to motivate asymptotic observer designs (Baras, Bensoussan, and James [2]; James [15]). Such designs were obtained by approximating the limiting filter, in the spirit of extended Kalman filtering, and required the system to satisfy a *detectability* condition. The purpose of this paper is to show that the approach of Baras and

*Received by the editors July 24, 1989; accepted for publication (in revised form) July 6, 1990. This research was supported in part by Systems Research Center, University of Maryland, under National Science Foundation grant CDR-85-00108 and AFOSR-87-0073, and by Lefschetz Center for Dynamical Systems, Brown University, under Army Research Office/Massachusetts Institute of Technology grant DAAL-03-86-K-0171.

†Department of Mathematics, University of Kentucky, Lexington, Kentucky 40506.

Krishnaprasad is successful in theory by carrying it out in two general settings: finite-state machines and nonlinear control systems.

The underlying asymptotic nonlinear filtering problem is connected with the theory of large deviations, which provides a link between asymptotic probabilities and variational problems. This problem has been studied by Hijab [12], [13], who characterized the limiting filter in terms of Mortensen's [24] method of deterministic minimum energy estimation. In the observer problem, the initial state is unknown, so we construct nonlinear filters for the randomized systems that reflect this lack of knowledge. If the system input is universal, then these filters are are consistent as $\epsilon \rightarrow 0$ (Theorems 3.1 and 4.1). The limiting filters so obtained are also consistent: they compute the state exactly (Theorems 3.2 and 4.2).

In §2 we review the concept of observability and define the term "finite time observer." In §3 we study in detail the method as it applies to finite-state machines. The observer obtained is finite-dimensional, and we describe how it can be implemented by a simplified parallel algorithm or circuit. These results might be useful for approximating more complex systems and developing numerical algorithms. In §4 we turn to continuous-time nonlinear control systems. The observers are, in general, infinite-dimensional (with dynamics given by Hamilton–Jacobi equations), a disadvantage shared with nonlinear stochastic filters. Because observability is so important, we study this concept further in §5 by characterizing the change in local structure of the sets of indistinguishable points as time increases, and relating this to certain properties of the value function, which solves the Hamilton-Jacobi equation. To do so, we introduce an observability grammian for nonlinear systems (a 2-form) and use it to define a time-dependent distribution, which we call the observability grammian distribution. This develops well-known results on local observability based on the observability codistribution introduced by Hermann and Krener [11]. We specialize our results to bilinear systems in §6.

2. Preliminaries. In this section we recall some definitions concerned with the concepts of observability and observers. We begin by introducing the models to be used.

We consider discrete-time finite-state machines defined by

$$\begin{aligned} &x(t+1) = f(x(t), u(t)); \quad t = 0, 1, 2, \cdots; \quad x(0) = x_0; \\ &y(t) = h(x(t)); \qquad t = 0, 1, 2, \cdots. \end{aligned} \tag{2.1}$$

The state $x(t)$ evolves in a finite set $\mathbf{X}$, and the control $u(t)$ and observation $y(t)$ take values in finite sets $\mathbf{U}$ and $\mathbf{Y}$, respectively. These sets have n, m, and p elements, respectively. The machine is described by a state transition map $f : \mathbf{X} \times \mathbf{U} \rightarrow \mathbf{X}$ and an output map $h : \mathbf{X} \rightarrow \mathbf{Y}$. Denote by $[0, T\,]$ the time interval $\{0, 1, 2, \cdots, T\}$ and define the sequence spaces $\mathcal{U}^T = \{u : [0, T\,] \rightarrow \mathbf{U}\}$, $\mathcal{U} = \bigcup_{T \geq 0} \mathcal{U}^T$, $\mathcal{X}^T = \{x : [0, T\,] \rightarrow \mathbf{X}\}$, $\mathcal{Y}^T = \{y : [0, T\,] \rightarrow \mathbf{Y}\}$. Let γ_u denote the flow for system (2.1); that is, $x(t) = \gamma_u(t)x_0 \in \mathbf{X}$ is the state at time t of (2.1) corresponding to the control $u \in \mathcal{U}$ with initial condition $x_0 \in \mathbf{X}$. We also write $x^T = (x(0), x(1), \cdots, x(T)) \in \mathcal{X}^T$ for an entire trajectory over the time interval $[0, T\,]$. Similarly, $y(t) = h(\gamma_u(t)x_0) \in \mathbf{Y}$ and $y^T = (y(0), y(1), \cdots, y(T)) \in \mathcal{Y}^T$.

The continuous-time nonlinear control systems we consider are defined by

$$\begin{aligned} &\dot{x}(t) = f(x(t), u(t)), \quad t > 0, \quad x(0) = x_0; \\ &y(t) = h(x(t)), \qquad t \geq 0. \end{aligned} \tag{2.2}$$

Here, $x(t) \in \mathbf{X}$, $u(t) \in \mathbf{U}$, and $y(t) \in \mathbf{Y}$; where $\mathbf{X} = \Re^n$, $\mathbf{Y} = \Re^p$ and $\mathbf{U} \subset \Re^m$. We assume that $f \in C^\infty(\Re^n \times \Re^m, \Re^n)$ and $h \in C^\infty(\Re^n, \Re^p)$ satisfy

$$\begin{aligned} |f(x,u) - f(z,u)| &\leq C|x-z|, \\ |f(x,u)|| &\leq C\,(1 + |x| + |u|)\,, \\ |h(x) - h(z)| &\leq C|x-z|, \end{aligned}$$

where $C > 0$ is independent of $x, z \in \Re^n$, $u \in \mathbf{U}$. Let $\mathcal{U}^T = C([0,T\,], \mathbf{U})$, $\mathcal{X}^T = C([0,T\,], \mathbf{X})$, $\mathcal{Y}^T = C([0,T\,], \mathbf{Y})$, denote the respective path spaces, equipped with the uniform norms, and write $\mathcal{U} = \bigcup_{T \geq 0} \mathcal{U}^T$. Our observer design (given in §4) is also valid for piecewise continuous controls, although we do not treat this extension here. The flow is again written $x(t) = \gamma_u(t)x_0$, etc.

In what follows, all time instants t and intervals $[0,T\,]$ are assumed to be within the domain of definition dom(u) of the relevant control $u \in \mathcal{U}$. In §3.2 and §4.2, the controls may be either open or closed loop. We turn now to the concept of observability. A system Σ will mean either (2.1) or (2.2). The definitions are the same for both.

DEFINITION 2.1 (a) A system Σ is observable on $[0,T\,]$ if for all $x_0^1 \neq x_0^2$ in $\mathbf{X}$ there exists $u \in \mathcal{U}$ such that

$$h(\gamma_u(t)x_0^1) \neq h(\gamma_u(t)x_0^2) \quad \text{for some} \quad t \in [0,T\,]. \tag{2.3}$$

(b) We say that Σ is reconstructable on $[0,T\,]$ if for all $x_0^1 \neq x_0^2$ in $\mathbf{X}$ there exists $u \in \mathcal{U}$ such that

$$\begin{aligned} h(\gamma_u(t)x_0^1) = h(\gamma_u(t)x_0^2) \quad &\text{for all} \quad t \in [0,T\,] \\ \text{implies} \quad \gamma_u(T)x_0^1 = \gamma_u(T)x_0^2&. \end{aligned} \tag{2.4}$$

(c) A control u for which (2.4) holds for all states $x_0^1 \neq x_0^2$ is termed *universal* on $[0,T\,]$.

For continuous-time systems, if u is universal on $[0,T\,]$, then (2.3) holds for all $x_0^1 \neq x_0^2$. The existence and genericity of universal controls is discussed by Sontag [26] and Sussmann [27]. A related discussion of observers is given by Sontag [26].

The observer problem is a deterministic state estimation problem: we wish to design a deterministic dynamical system that takes as inputs $u \in \mathcal{U}^t$ and $y \in \mathcal{Y}^t$ and produces an estimate $\hat{x}(t) \in \mathbf{X}$ of the state $x(t) \in \mathbf{X}$; the initial condition $x_0 \in \mathbf{X}$ being unknown. This problem was first solved for observable linear systems by Luenberger [22], in an asymptotic sense: $\lim_{t\to\infty} \operatorname{dist}(\hat{x}(t), x(t)) = 0$. Later, Wonham extended this result to *detectable* linear systems [31], which are not necessarily observable. Asymptotic observers for nonlinear systems have been obtained by numerous authors, for instance: Williamson [30], Kuo, Elliott, and Tarn [21], Bestle and Zeitz [4], Krener and Respondek [20], Isidori [14], Baras, Bensoussan, and James [2], Celle et al. [6]. These observers are defined by ordinary differential equations, are finite-dimensional, and either assume a local observability condition or involve a stability condition for the equation satisfied by the error (detectability). Each design has a number of advantages and disadvantages (Walcott, Corless, and Zak [29]), and many require precise knowledge of the system model.

In this paper we introduce observers that compute the state *exactly* after a finite time has elapsed:

$$\hat{x}(t) \;=\; x(t) \quad \text{for all} \quad t \geq T. \tag{2.5}$$

An important requirement is that the estimate be computed recursively; thus we want the observer to be realizable as a dynamical system, of the general form

$$\begin{aligned} \dot{m}(t) &= F(m(t), u(t), y(t)), \quad t > 0, \quad m(0) = m_0; \\ \hat{x}(t) &= G(m(t)), \qquad t \geq 0; \end{aligned} \tag{2.6}$$

for continuous-time systems, and

$$\begin{cases} m(t+1) = F(m(t), u(t), y(t)), \quad t = 0, 1, 2, \cdots; \quad m(0) = m_0; \\ \hat{x}(t) = G(m(t)), \qquad t = 0, 1, 2, \cdots; \end{cases} \tag{2.7}$$

for discrete-time systems. Equations (2.6) and (2.7) are interpreted as defining dynamical systems with state $m(t)$ taking values in a space $\mathbf{M}$, and producing an output $\hat{x}(t) \in \mathbf{X}$. This output is the state estimate. The space $\mathbf{M}$ need not be finite-dimensional.

DEFINITION 2.2 A dynamical system $\mathcal{O}_\Sigma$ of the form (2.6) or (2.7) is called a finite time observer for the system Σ provided (2.5) holds for any control $u \in \mathcal{U}$ that is universal on $[0, T]$.

The task of observer design, then, is to somehow obtain from the given data a system of the form (2.6) or (2.7) that fulfills the requirements of the definition. In the following we achieve this goal in some generality using ideas from nonlinear filtering and large deviations.

3. Finite-state machines.

3.1. Asymptotic filtering. We regard system (2.1) as a deterministic Markov chain and define a random perturbation as follows. Let $N(x)$ denote a set of points "neighboring" $x \in \mathbf{X}$, defined so that $N(x)$ contains $f(x, u)$ for every $u \in \mathbf{U}$, and every point z for which $x = f(z, u)$ for some $u \in \mathbf{U}$. Define for $z, x \in \mathbf{X}$ and $u \in \mathbf{U}$

$$U(z, x; u) = \begin{cases} 0 & \text{if} \quad x = f(z, u), \\ 1 & \text{if} \quad x \neq f(z, u) \quad \text{and} \quad x \in N(z), \\ +\infty & \text{if} \quad x \neq f(z, u) \quad \text{and} \quad x \notin N(z), \end{cases}$$

and for $\epsilon > 0$,

$$A^\epsilon(u)_{xz} = \frac{1}{Z} \exp\left(-\frac{1}{\epsilon} U(z, x; u)\right),$$

where Z is a normalization constant: $\sum_{x \in N(z)} A^\epsilon(u)_{xz} = 1$ for each $z \in \mathbf{X}$. (Throughout, Z denotes an appropriate normalization constant.) A random perturbation of the state $\{x(t);\ t = 0, 1, 2, \cdots\}$ is an $\mathbf{X}$-valued Markov chain $\{x^\epsilon(t);\ t = 0, 1, 2, \cdots\}$ whose probabilities $p^\epsilon_x(t) = \text{Prob}\,(x^\epsilon(t) = x \mid u, x^\epsilon(0) = x_0)$ evolve according to

$$p^\epsilon(t+1) = A^\epsilon(u(t)) p^\epsilon(t); \quad t = 0, 1, 2, \cdots; \quad p^\epsilon(0) = p_0. \tag{3.1}$$

Note that $A^\epsilon(u) \to A(u)$ and thus $x^\epsilon(t) \Rightarrow x(t)$ as $\epsilon \to 0$.

Denoting by θ an element of $\mathcal{X}^T$, we have, using the Markov property,

$$\begin{aligned} P^\epsilon_{u, x_0}(\theta) &= \text{Prob}\,(x^{\epsilon, T} = \theta \mid u, x^\epsilon(0) = x_0), \\ &= \frac{1}{Z} \exp\left(-\frac{1}{\epsilon} \sum_{t=0}^{T-1} U(\theta(t), \theta(t+1); u(t))\right) I_{\{\theta(0) = x_0\}}. \end{aligned} \tag{3.2}$$

This is a probability measure on $\mathcal{X}^T$ corresponding to the process $x^{\epsilon, T}$ (a Gibbs distribution). From this explicit formula, we readily obtain a large deviation result for these measures (cf. Theorem 3.1 below).

The noisy observation is a $\mathbf{Y}$-valued process $\{y^\epsilon(t);\ t = 0, 1, 2, \cdots\}$ distributed according to the conditional probability

$$\text{Prob}\,(y^\epsilon(t) = y \mid x^\epsilon(t) = x) = \frac{1}{Z} \exp\left(-\frac{1}{\epsilon} V(x, y)\right), \tag{3.3}$$

where

$$V(x, y) = \begin{cases} 0 & \text{if} \quad y = h(x), \\ 1 & \text{if} \quad y \neq h(x). \end{cases}$$

Let us now define a filter, assuming instead that the initial condition x_0^ϵ is random and has the uniform distribution μ on $\mathbf{X}$. This is done because in the observer problem the initial condition is completely unknown. For $\theta \in \mathcal{X}^T$ and $\eta \in \mathcal{Y}^T$, Bayes' rule gives

$$\begin{aligned} \Pi^\epsilon_{u,\mu}(\theta \mid \eta) &= \text{Prob}\,(x^{\epsilon,T} = \theta \mid y^{\epsilon,T} = \eta, u, \mu) \\ &= \frac{1}{Z} \exp\left(-\frac{1}{\epsilon}\left[\sum_{t=0}^{T-1} U(\theta(t), \theta(t+1); u(t)) + \sum_{t=0}^{T} V(\theta(t), \eta(t))\right]\right), \end{aligned} \tag{3.4}$$

where the averaging with respect to μ is absorbed into the normalization constant Z.

Let x_0^* denote the actual initial condition of (2.1), and let x_*^ϵ, x_*, y_*^ϵ, y_* denote the corresponding state and observation trajectories produced by (2.1) and its random perturbation, under the action of a control $u \in \mathcal{U}$. We are interested in the asymptotic behaviour of this filter evaluated using the actual noisy observations y_*^ϵ. Define the *action function* for the filters by

$$I_u(\theta \mid \eta) = \sum_{t=0}^{T-1} U(\theta(t), \theta(t+1); u(t)) + \sum_{t=0}^{T} V(\theta(t), \eta(t)). \tag{3.5}$$

We then have the following large deviations and consistency result.

THEOREM 3.1. *The family of filters* $\{\Pi^\epsilon_{u,\mu}\}$ *obey the large deviation principle in probability* (LDPP) *with action function* $I_u(\theta \mid y_*)$. *In particular, if* $A \subset \mathcal{X}^T$, *then*

$$\lim_{\epsilon \to 0} \epsilon \log \Pi^\epsilon_{u,\mu}(A \mid y_*^\epsilon) = -\min_{\theta \in A} I_u(\theta \mid y_*) \quad \textit{in probability.}$$

Furthermore, if $u \in \mathcal{U}^T$ *is a control for which* (2.3) *holds for all* $x_0^1 \neq x_0^2$, *then*

$$\Pi^\epsilon_{u,\mu}(\cdot \mid y_*^\epsilon) \Rightarrow \delta_{x_*} \quad \textit{in probability as } \epsilon \to 0.$$

The proof of this theorem is straightforward and employs the explicit formula (3.4). The weak convergence statement depends on the fact that the assumed observability condition implies that the functional $I_u(\cdot \mid y_*)$ has a *unique* minimizer, namely, x_*. Indeed, $I(x_* \mid y_*) = 0$, and also if $I(\theta \mid y_*) = 0$, then θ is a solution of (2.1) producing the output y_*. If for the given control (2.3) holds for all $x_0^1 \neq x_0^2$, then this implies that $\theta(0) = x_*(0)$, and hence $\theta = x_*$ (cf. Theorem 3.2 below). Thus if A is a subset of $\mathcal{X}^T$ not containing x_*, then the large deviation limit result implies (Varadhan [28]) that $\Pi^\epsilon_{u,\mu}(A \mid y_*^\epsilon)$ decays exponentially to zero as $\epsilon \to 0$. This theorem is interpreted as saying that the filters $\Pi^\epsilon_{u,\mu}$, when applied to the actual noisy observations, concentrate on the desired state trajectory x_* as $\epsilon \to 0$, provided the observability condition holds.

3.2. Observer design. It is clear from the above that the state trajectory can be determined by solving a variational problem: find the unique minimizer of $I_u(\cdot \mid y_*)$. We solve this problem using the dynamic programming method, and obtain a recursive system that becomes our finite time observer for the machine (2.1). For each $t = 0, 1, 2, \cdots$ define a function

$$m(\cdot, t) : \mathbf{X} \to [0, \infty)$$

by

$$m(x,t) = \min \{ I_u(\theta \mid y_*) : \theta(t) = x, \theta \in \mathcal{X}^t \} . \tag{3.6}$$

The function $I_u(\cdot \mid y_*)$ is taken to be defined on the interval $\{0, 1, \cdots, t\}$ rather than the interval $\{0, 1, \cdots, T\}$. The dynamic programming method gives

$$\begin{aligned} m(x,t) &= \min_{z \in N(x)} \{ m(z, t-1) + U(z, x; u(t-1)) + V(x, y_*(t)) \} ; \ t \geq 1; \\ m(x, 0) &= V(x, y_*(0)). \end{aligned} \tag{3.7}$$

Define also the "deterministic estimate" [24], [12]:

$$\hat{x}(t) = \operatorname*{argmin}_{x \in \mathbf{X}} m(x,t) \equiv \{ x \in \mathbf{X} : m(x,t) = 0 \} . \tag{3.8}$$

This quantity is set-valued, since $m(\cdot, t)$ may have many minima. For example, $\hat{x}(0) = \{ x \in \mathbf{X} : h(x) = y_*(0) \}$. Equations (3.7) and (3.8) define the limiting filter; in fact, $\lim_{\epsilon \to 0} \epsilon \log P(x^\epsilon(t) = x \mid y_*^{\epsilon,t}, u, \mu) = -m(x,t)$ in probability. Our main result concerning observers for the finite-state machine (2.1) is the following theorem.

THEOREM 3.2. *The recursive system* (3.7), (3.8) *is a* finite time observer *for the finite-state machine* (2.1). *In particular, if* u *is a* universal *control on* $[0, T]$, *then*

$$\hat{x}(t) = \{ x_*(t) \} \quad \textit{for all} \quad t \geq T. \tag{3.9}$$

Proof. For all $t = 0, 1, 2, \cdots$, $m(x_*(t), t) = 0$ and $m(x,t) \geq 0$ for all $x \in \mathbf{X}$. Suppose that $t \geq T$ and $\tilde{x} \in \mathbf{X}$ is such that $m(\tilde{x}, t) = 0$. Then there exists $\tilde{\theta} \in \mathcal{X}^t$ such that $\tilde{\theta}(t) = \tilde{x}$ and $I_u(\tilde{\theta} \mid y_*) = 0$. This implies that

$$\begin{aligned} \tilde{\theta}(s+1) &= f(\tilde{\theta}(s), u(s)); \qquad s = 0, 1, \cdots, T-1; \\ y_*(s) &= h(\tilde{\theta}(s)); \qquad s = 0, 1, \cdots, T. \end{aligned}$$

(In fact, these relations also hold for $s = T, \cdots, t-1$; $s = T+1, \cdots, t$, respectively.) If u is a universal control on $[0, T]$, this forces $\tilde{\theta}(T) = x_*(T)$. From this, it follows that $\tilde{\theta}(s) = x_*(s)$, $s = T+1, \cdots, t$, and so $\tilde{x} = \tilde{\theta}(t) = x_*(t)$. □

Computing $\hat{x}(t)$ requires the determination of those $x \in \mathbf{X}$ satisfying $m(x,t) = 0$. For each x, the computation of $m(x,t)$ involves up to n minimizations. This procedure can be simplified by setting

$$\tilde{m}(x,t) = \begin{cases} 0 & \text{if } m(x,t) = 0, \\ 1 & \text{if } m(x,t) > 0. \end{cases}$$

Regard 1 and 0 as the logical values TRUE and FALSE. Then $\tilde{m}(x,t)$ is given by the logical expression

$$\begin{aligned} \tilde{m}(x,t) &= V(x, y_*(t)) \vee \bigwedge_{z \in N(x) : x = f(z, u(t-1))} \tilde{m}(z, t-1) \quad \text{if } t \geq 1, \\ \tilde{m}(x, 0) &= V(x, y_*(0)), \end{aligned} \tag{3.10}$$

where the symbols $\vee$ and $\wedge$ denote OR and AND, respectively. If the set $\{ z \in N(x) : x = f(z, u(t-1)) \}$ is empty, we set $\tilde{m}(x,t) = 1$. Note that $\hat{x}(t) = \{ x : \tilde{m}(x,t) = 0 \}$. A digital circuit or parallel algorithm can readily be designed to realize (3.10).

4. Nonlinear control systems.

4.1. Asymptotic filtering. On a probability space $(\Omega, \mathcal{F}, P)$ we consider a family of diffusion processes $\{x^\epsilon(t), t \geq 0\}$ together with observation processes $\{\xi^\epsilon(t), t \geq 0\}$, satisfying the stochastic differential equations

$$\begin{aligned} dx^\epsilon(t) &= f(x^\epsilon(t), u(t))dt + \sqrt{\epsilon}dw(t), \qquad x^\epsilon(0) = x_0, \\ d\xi^\epsilon(t) &= h(x^\epsilon(t))dt + \sqrt{\epsilon}dv(t), \qquad \xi^\epsilon(0) = 0. \end{aligned} \tag{4.1}$$

Here, $\{w(t), t \geq 0\}$ and $\{v(t), t \geq 0\}$ are independent Wiener processes in $\mathbf{X} = \Re^n$ and $\mathbf{Y} = \Re^p$, respectively, and $x_0 \in \Re^n$. This defines the perturbed state and observation equations for the continuous-time nonlinear control system (2.2). As is well known (Freidlin and Wentzell [10]), $x^{\epsilon,T} \to x^T$ and $\xi^{\epsilon,T} \to \xi^T$ in probability as $\epsilon \to 0$, where $\xi(t) = \int_0^t y(s)ds$, and the distributions P^ϵ_{u,x_0} of $x^{\epsilon,T}$ on $\mathcal{X}^T$ obey the large deviation principle (uniformly in $x_0 \in \Re^n$).

We use the well-known robust filter (Davis [8]) for (4.1), defined assuming that the distribution of the initial condition x^ϵ_0 is a probability measure μ on $\Re^n$ such that $\mu(B(x,r)) > 0$ for every ball $B(x,r)$. This filter is a continuous map

$$\Pi^\epsilon_{u,\mu} : \mathcal{Y}^T \to \mathcal{P}(\mathcal{X}^T)$$

such that

$$\Pi^\epsilon_{u,\mu}(A \mid \xi^\epsilon_*) = P\left(x^{\epsilon,T} \in A \mid \xi^\epsilon_*,\ u, \mu\right) \ \text{ a.s.} \qquad (A \in \mathcal{B}(\mathcal{X}^T)).$$

An explicit formula for the filter corresponding to a bilinear system is given in §6. For $\theta \in \mathcal{X}^T$ and absolutely continuous $\eta \in \mathcal{Y}^T$, define the action function for these filters by

$$I_u(\theta \mid \eta) = \begin{cases} \frac{1}{2}\int_0^T |\dot\theta(s) - f(\theta(s), u(s))|^2 + |\dot\eta(s) - h(\theta(s))|^2 ds \\ \qquad\qquad\qquad\qquad \text{if } \theta \text{ is absolutely continuous,} \\ +\infty \qquad\qquad\qquad \text{otherwise.} \end{cases} \tag{4.2}$$

The $(*)$ notation of the preceding section is again used to indicate a distinguished initial condition and the corresponding state and observation paths. Parallel to Theorem 3.1 we have the following result, which refines the large deviation results of Hijab [12], [13], and James [15]; see also [18], [17].

THEOREM 4.1. *The filters* $\{\Pi^\epsilon_{u,\mu}\}$ *obey the* LDPP *with action function* $I_u(\theta \mid \xi_*)$. *In particular, if* $A \subset \mathcal{X}^T$, *then*

$$\Pi^\epsilon_{u,\mu}(A \mid \xi^\epsilon_*) \asymp \exp\left(-\frac{1}{\epsilon}\inf_{\theta \in A} I_u(\theta \mid \xi_*)\right) \quad \textit{in probability as} \quad \epsilon \to 0.$$

Furthermore, if u is a universal *control on* $[0, T]$, *then*

$$\Pi^\epsilon_{u,\mu}(\cdot \mid \xi^\epsilon_*) \Rightarrow \delta_{x_*} \ \textit{ in probability as } \epsilon \to 0.$$

Here, the symbol $\asymp$ denotes asymptotic equivalence. The proof of this theorem is omitted and is based on a generalization of the well-known Varadhan–Laplace asymptotic method (Varadhan [28]), and the fact that if u is universal on $[0, T]$, then the unique minimizer of $I_u(\cdot \mid \xi_*)$ is x_* (cf. Theorem 4.2 below). The convergence of the measures is with respect to the Prohorov metric on the space $\mathcal{P}(\mathcal{X}^T)$ of probability measures.

4.2. Observer design. As for finite-state machines, the state trajectory can be obtained in principle by finding the minimizer of $I_u(\cdot \mid \xi_*)$. For $t \geq 0$ define a function

$$m(\cdot, t) : \Re^n \to [0, \infty)$$

by

$$m(x,t) = \inf \left\{ \tfrac{1}{2}|h(\theta(0)) - y_*(0)|^2 + I_u(\theta \mid \xi_*) : \theta(t) = x, \theta \in \mathcal{X}^t \right\}, \tag{4.3}$$

where $I_u(\cdot \mid \xi_*)$ is defined for the time interval $[0,t]$. The cost term for the initial condition does not appear in the limiting filter. It is added here to include the information that is available at time $t = 0$ in the deterministic limit. The value function $m(x,t)$ is continuous and is the unique viscosity solution of the Hamilton–Jacobi (HJ) equation (Crandall, Evans, and Lions [7])

$$\begin{cases} \dfrac{\partial}{\partial t} m(x,t) + \frac{1}{2}|Dm(x,t)|^2 + f(x,u(t)) \cdot Dm(x,t) \\ \qquad\qquad\qquad -\frac{1}{2}|y_*(t) - h(x)|^2 = 0 \quad \text{in } \Re^n \times (0,\infty), \\ \qquad m(x,0) = \frac{1}{2}|y_*(0) - h(x)|^2 \qquad \text{in } \Re^n. \end{cases} \tag{4.4}$$

Also define

$$\hat{x}(t) = \operatorname*{argmin}_{x \in \Re^n} m(x,t). \tag{4.5}$$

The limiting filter satisfies (4.4), (4.5), except with $m(\cdot, 0) = 0$; if $\pi^\epsilon_{u,\mu}(x, t \mid \xi^{\epsilon,t}_*)$ denotes the conditional density, then $\lim_{\epsilon \to 0} \epsilon \log \pi^\epsilon_{u,\mu}(x,t \mid \xi^{\epsilon,t}_*) = -m(x,t)$ in probability [17], [15]. We have the following result concerning the existence of observers for the nonlinear system (2.2).

THEOREM 4.2. *The dynamical system* (4.4), (4.5) *is a* finite time observer *for the nonlinear system* (2.2). *In particular, if u is a* universal *control on* $[0, T]$, *then*

$$\hat{x}(t) = \{x_*(t)\} \quad \textit{for all} \quad t \geq T. \tag{4.6}$$

Proof. Now $m(x,t) \geq 0$ for every $x \in \Re^n$, and also $m(x_*(t), t) = 0$ for $t \geq 0$. Let $t \geq T$ and suppose that u is universal on $[0, T]$. If $z \in \Re^n$ is such that $m(z,t) = 0$, then, thanks to the lower semicontinuity of $I_u(\cdot \mid \xi_*)$, there exists $\theta \in \mathcal{X}^t$ such that $\frac{1}{2}|h(\theta(0)) - y_*(0)|^2 + I_u(\theta \mid \xi_*) = 0$ and $\theta(t) = z$. Hence $\dot{\theta}(s) = f(\theta(s), u(s))$ and $y_*(s) = h(\theta(s))$ for each $s \in [0,t]$. Thus θ is a state trajectory, and produces the output y_*. Since u is universal on $[0, T]$, we must have $\theta(T) = x_*(T)$, and therefore $z = x_*(t)$. □

Remarks. (i) Note that this observer is *infinite-dimensional*, with state space $\mathbf{M} = C(\Re^n)$.

(ii) Theorem 4.2 does not require all the smoothness assumed for the data f and h in §2. The Lipschitz continuity and growth conditions suffice.

(iii) What is important is the computation of $\hat{x}(t) = \{x \in \Re^n : m(x,t) = 0\}$, and, as noted in §3.2, it may be possible numerically to compute $\hat{x}(t)$ without computing the complete solution of (4.4) (James [16]).

5. Observability and the value function. In this section we study the time evolution of the sets of indistinguishable points and certain properties of the value function $m(x,t)$. This provides some insight into how an observer can aquire information and determine the state trajectory. We consider system (2.2) and drop the $(*)$ notation used in previous sections. Most of the results below remain valid if the class of admissible controls $\mathcal{U}$ is expanded to include all measurable controls $u : [0, T] \to \mathbf{U}$.

For each $t \geq 0$, define the set of points *indistinguishable* from x_0 on $[0,t]$ with respect to the control $u \in \mathcal{U}^t$ by

$$\mathbf{I}_t^u(x_0) = \{x \in \Re^n : h(\gamma_u(s)x) = h(\gamma_u(s)x_0) \text{ for all } 0 \leq s \leq t\}, \tag{5.1}$$

and the set of points indistinguishable from x_0 on $[0,t]$ with respect to the class of controls $\mathcal{U}^t$ by

$$\mathbf{I}_t(x_0) = \bigcap_{u \in \mathcal{U}^t} \mathbf{I}_t^u(x_0). \tag{5.2}$$

As time t increases, these sets can decrease in both size and dimension. For instance, $\mathbf{I}_0(x_0)$ is generically the submanifold $h^{-1}(h(x_0))$, whereas if the control system is observable on $[0,T]$, then, for all $t \geq T$, $\mathbf{I}_t(x_0) = \{x_0\}$. Below, we study the local structure of the sets $\mathbf{I}_t(x_0)$ for times t between these extremes.

We define the *observability grammian* $\mathcal{O}^u$ for system (2.2) given the control $u \in \mathcal{U}$ to be a map assigning to each $t \in \operatorname{dom}(u)$ a field $\mathcal{O}^u(t)$ of symmetric bilinear forms defined for $X_1, X_2 \in T_x\Re^n$ by

$$\begin{aligned}\mathcal{O}_x^u(t)(X_1,X_2) &= \langle dh(x), X_1\rangle \cdot \langle dh(x), X_2\rangle \\ &\quad + \int_0^t \langle dh(\gamma_u(s)x), \gamma_{u*x}(s)X_1\rangle \cdot \langle dh(\gamma_u(s)x), \gamma_{u*x}(s)X_2\rangle ds.\end{aligned} \tag{5.3}$$

Here, $\gamma_{u*x}(s)$ denotes the differential of the flow, $\langle\cdot,\cdot\rangle$ denotes the pairing between vectors and covectors, and the dot is the usual inner product in $\Re^p$.

Remark. This grammian is "dual" to the deterministic Malliavin covariance matrix defined by Bismut [5]. The observability grammian for linear systems was introduced by Kalman [19].

The rank of $\mathcal{O}_x^u(t)$ is a left-continuous nondecreasing function of time t for each fixed $x \in \Re^n$. The kernel of this grammian is defined by

$$\Delta_t^u(x) = \ker \mathcal{O}_x^u(t) \tag{5.4}$$

for $x \in \Re^n$. Define the *observability grammian distribution* Δ_t for $t \geq 0$, $x \in \Re^n$ by

$$\Delta_t(x) = \bigcap_{u \in \mathcal{U}^t} \Delta_t^u(x). \tag{5.5}$$

The set $\mathcal{M}_t$ of nonsingular points of Δ_t is an open dense subset of $\Re^n$. This distribution characterizes the local structure of the sets of indistinguishable points.

THEOREM 5.1. *Fix $t \geq 0$ and let $x_0 \in \Re^n$ be a nonsingular point for Δ_t,; $d = \dim \Delta_t(x_0)$. Then there exists a neighborhood U of x_0 such that $\mathbf{I}_t(x_0) \cap U$ is a d-dimensional integral submanifold of Δ_t.*

Proof. Since $\dim \Delta_t$ is constant near x_0, and the $\mathcal{O}_x^u(t)$ are continuous functions of x, there exists controls $u_1, \cdots, u_\ell \in \mathcal{U}^t$ such that

$$\Delta_t(x) = \bigcap_{i=1}^{\ell} \Delta_t^{u_i}(x)$$

for all x near x_0. Write $H = L^2([0,t], \Re^p)$ and define $Y_t : \Re^n \to H^\ell$ by

$$Y_t(x) = \begin{Bmatrix} Y_t^1(x_0) \\ \vdots \\ Y_t^\ell(x_0) \end{Bmatrix},$$

where $Y_t^i(x) = \{h(\gamma_{u_i}(s)x),\ 0 \le s \le t\}$. Now $dY_t^i(x) = \{\langle dh(\gamma_{u_i}(s)x), \gamma_{u_i *}(s)\cdot\rangle,\ 0 \le s \le t\}$, and so

$$\ker dY_t(x) = \Delta_t(x)$$

for all x near x_0. By the rank theorem (Rudin [25]), there is a neighborhood V of x_0 such that $Y_t(V)$ is an $(n-d)$-dimensional submanifold of H^ℓ. By the implicit function theorem, there exists a neighborhood U of x_0 contained in V such that $Y_t^{-1}(Y_t(x_0))\cap U$ is a d-dimensional submanifold of $\Re^n$. This submanifold is an integral submanifold of Δ_t through x_0. By construction, $\mathbf{I}_t(x_0) \cap U \subset Y_t^{-1}(Y_t(x_0)) \cap U$.

Now let $u \in \mathcal{U}^t$, and $x \in Y_t^{-1}(Y_t(x_0)) \cap U$. Reducing U if necessary, there exists a piecewise smooth curve $\alpha : [0,1] \to Y_t^{-1}(Y_t(x_0)) \cap U$ such that $\alpha(0) = x_0$, $\alpha(1) = x$ and

$$\dot\alpha(r) \in \Delta_t(\alpha(r)) \subset \Delta_t^u(\alpha(r)) \ \text{ a.e. } r \in [0,1].$$

This implies

$$\langle dh(\gamma_u(s)\alpha(r)), \gamma_{u*}(s)\dot\alpha(r)\rangle = 0$$

for $0 \le s \le t$, almost everywhere $r \in [0,1]$. Then $h(\gamma_u(s)\alpha(r)) = h(\gamma_u(s)x_0)$ for all $0 \le s \le t$, $r \in [0,1]$, and hence $h(\gamma_u(s)x) = h(\gamma_u(s)x_0)$ for all $0 \le s \le t$. Therefore $Y_t^{-1}(Y_t(x_0)) \cap U \subset \mathbf{I}_t(x_0) \cap U$, completing the proof. □

Let us say that system (2.2) is *locally observable* on $[0,T\,]$ if every point $x_0 \in \Re^n$ has a neighborhood U_{x_0} such that $I_T(x_0) \cap U_{x_0} = \{x_0\}$.

COROLLARY 5.1. *If* $\Delta_T \equiv \{0\}$ *in* $\Re^n$, *then system* (2.2) *is locally observable on* $[0,T\,]$. *Conversely, if* (2.2) *is* (*locally*) *observable on* $[0,T\,]$, *then* $\Delta_T = \{0\}$ *generically.*

Proof. The first statement follows immediately from Theorem 5.1. The dimension of Δ_T is constant on the open dense set $\mathcal{M}_T$, and Theorem 5.1 implies that this dimension must be zero there if (2.2) is locally observable on $[0,T\,]$; hence the second claim. □

In [11], Hermann and Krener introduced the *observability codistribution* Ω defined for $x \in \Re^n$ by

$$\Omega(x) = \operatorname{span}_\Re \left\{dh(x), L_{f^u} dh(x), L_{f^u}^2 dh(x), \cdots \ :\ u \in \mathbf{U}\right\},$$

where $L_{f^u}^k dh$, $k = 0,1,2,\cdots$, denotes the kth Lie derivative of the 1-form dh with respect to the vector field with constant control $f^u = f(\cdot,u)$. The annihilator of Ω is denoted $\Omega^\perp$.

PROPOSITION 5.1. *For all* $t > 0$ *we have*

$$\Delta_t(x) \subset \Omega^\perp(x) \ \text{ for all } x \in \Re^n.$$

Furthermore, if $\mathcal{R}$ *denotes the open dense subset of nonsingular points for* Ω *and if* $\mathbf{U}$ *is compact, then for each* $x_0 \in \mathcal{R}$ *there exists* $t_0 > 0$ *and a neighborhood* $V_{x_0} \subset \mathcal{R}$ *of* x_0 *such that*

$$\Delta_t(x) = \Omega^\perp(x) \ \text{ for all } x \in V_{x_0}, \ \text{ and all } t \in (0,t_0].$$

Proof. 1. Let $X \in \Delta_t(x)$ and let $u(s) = u \in \mathbf{U}$ for $s \in [0,t]$. Then $X \in \Delta_t^u(x)$ and so $\langle dh(\gamma_u(s)x), \gamma_{u*x}(s)X\rangle = 0$, $0 \le s \le t$. Repeatedly differentiating with respect to s gives for $k = 0,1,2,\cdots$ $\langle L_{f^u}^k dh(x), X\rangle = 0$. This holds for each fixed $u \in \mathbf{U}$, and hence $X \in \Omega^\perp(x)$.

2. Since $\mathbf{U}$ is compact, there exits $t_0 > 0$ and a neighborhood $V_{x_0} \subset \mathcal{R}$ of x_0 such that if $x \in V_{x_0}$ and $t \in [0,t_0]$ then $\gamma_u(s)x \in \mathcal{R}$, $0 \le s \le t$, for any $u \in \mathcal{U}^t$. Let $X \in \Omega^\perp(x)$, $x \in V_{x_0}$, and $0 < t \le t_0$. By approximation, it is enough to show that if $u \in \mathcal{U}^t$ is piecewise constant, then $X \in \Delta_t^u(x)$. Select points $t^0 = 0 < t^1 < \cdots < t^k = t$ and

$u^i \in \mathbf{U}$ and define $u(s) = u^i$ if $s \in [t^i, t^{i-1})$, $i = 1, \cdots, k$. The invariance of $\Omega^\perp$ under the action of f^{u^i} (Isidori [14]) implies that $\gamma_{u^i * x^{i-1}}(s)\Omega^\perp(x^{i-1}) \subset \Omega^\perp(\gamma_{u^i}(s)x^{i-1})$ for $s \in [t^i, t^{i-1}]$, where $x^i = \gamma_{u^i}(t^i)x^{i-1}$, $x^0 = x$. Then $\langle dh(\gamma_u(s)x), \gamma_{u*x}(s)X\rangle = 0$ for $0 \leq s \leq t$; i.e., $X \in \Delta_t^u(x)$. □

Thus, generically, the leaves of Ω consist of points that are not *instantaneously* distinguishable. System (2.2) satisfies the *observability rank condition* (ORC), provided that $\dim \Omega(x) = n$ for all $x \in \Re^n$. Hermann and Krener prove that if the control system satisfies the ORC, then it is locally weakly observable; i.e., every point can instantaneously be distinguished from it neighbors; conversely, if this property holds, then the ORC holds generically. Note that if the ORC holds, then necessarily $\Delta_t \equiv \{0\}$ in $\Re^n$ for all $t > 0$. Also, note that $\Omega^\perp(x) \subset \Delta_0(x)$ for all $x \in \Re^n$.

A control $u \in \mathcal{U}^T$ is called a *local universal control* on $[0, T]$ if every $x_0 \in \Re^n$ has a neighborhood V_{x_0} such that $\mathbf{I}_T^u(x_0) \cap V_{x_0} = \{x_0\}$. The observability grammian $\mathcal{O}^u$ is related to the system obtained by linearizing (2.2) along a trajectory.

COROLLARY 5.2. *If u is a local universal control on $[0, T]$, then the linearized system*

$$\begin{cases} \dot{X}(t) = Df(\gamma_u(t)x, u(t))X(t), & t > 0, \quad X(0) = X_0 \in T_x\Re^n; \\ Y(t) = \langle dh(\gamma_u(s)x), X(t)\rangle, & t \geq 0 \end{cases} \tag{5.6}$$

is observable on $[0, T]$ for all x belonging to an open dense subset of $\Re^n$. Conversely, if (5.6) is observable on $[0, T]$ for every $x \in \Re^n$, then u is a local universal control.

Proof. Specializing Theorem 5.1 to the case of a single control, it follows that if u is a (local) universal control on $[0, T]$, then Δ_T^u is zero on an open dense subset of $\Re^n$; that is, the observability grammian $\mathcal{O}_x^u(T)$ is strictly positive definite for all x in this set. This implies that the linear system (5.6) is observable on $[0, T]$, for all such x. The converse statement also follows from Theorem 5.1. □

These results give information on the sets of points $\hat{x}(t)$ on which the value function $m(\cdot, t)$ vanishes.

THEOREM 5.2. *Let $u \in \mathcal{U}$. Then we have:w*

(a) *There exists a neighborhood $S(x_0)$ of $\Gamma(x_0) = \{\gamma_u(t)x_0;\ t \geq 0\}$ on which the value function m is of class $C^{\infty,1}$, and $P(t) = D^2m(\gamma_u(t)x_0, t)$ satisfies the Riccati equation*

$$\begin{cases} \dot{P}(t) = -P(t)Df(\gamma_u(t)x_0, u(t)) - Df(\gamma_u(t)x_0, u(t))'P(t) - P(t)^2 \\ \qquad\qquad + Dh(\gamma_u(t)x_0)'Dh(\gamma_u(t)x_0), \qquad t > 0; \\ P(0) = Dh(x_0)'Dh(x_0); \end{cases} \tag{5.7}$$

(b) *If x_0 is a nonsingular point for Δ_t^u, then there exists a neighborhood W of $\gamma_u(t)x_0$ such that $\hat{x}(t; x_0, u) \cap W$ is a submanifold of dimension $\dim \Delta_t^u(x_0)$;*

(c) *If u is a (local) universal control on $[0, T]$, then for $t \geq T$ we have*

$$D^2m(\gamma_u(t)x_0, t) > 0, \tag{5.8}$$

for all x_0 belonging to an open dense subset of $\Re^n$.

Proof. Part (a). For any $t \geq 0$, there exists a minimizer $\tilde{\theta} \in \mathcal{X}^t$ such that $\tilde{\theta}(t) = \gamma(t)x_0$, $m(\gamma(t)x_0, t) = 0$. This implies that $\tilde{\theta}(s) = \gamma(s)x_0$, $0 \leq s \leq t$, and so the minimizer is unique. Therefore $m(\cdot, \cdot)$ is differentiable at $(\gamma(t)x_0, t)$ and it also follows that $(\gamma(t)x_0, t)$ is not conjugate to $\Gamma(x_0)$ (Fleming [9]). The method of characteristics then provides a unique smooth solution $\tilde{m}$ to the HJ equation (9.2) in a neighborhood $S(x_0)$ of $\Gamma(x_0)$; this solution coincides with m in $S(x_0)$.

The Riccati equation (5.7) is obtained by differentiating $D^2m(\gamma(s)x_0, s)$ with respect to s and combining the resulting expression with the equation obtained by differentiating the HJ equation (4.4) twice with respect to x. Note that the adjoint variable $\lambda(s) = Dm(\gamma(s)x_0, s)$ equals zero and $y(s) = h(\gamma(s)x_0)$ on $[0, t]$.

Part (b). We have $\hat{x}(t; x_0, u) = \gamma_u(t)\mathbf{I}_t^u(x_0)$, and so the result of Theorem 5.1 specialized to a single control can be carried over by the flow to establish (b).

Part (c). Corollary 5.2 implies that if u is a (local) universal control on $[0, T\,]$, then for $t \geq T$ we have $\mathcal{O}^u_{x_0}(t) > 0$ for x_0 belonging to an open dense subset of $\Re^n$. From linear systems theory (Kalman [19]), if $\mathcal{O}^u_{x_0}(T) > 0$ then the solution of the Riccati equation $P(t)$ is also strictly positive definite for $t \geq T$. □

Remark. In the context of stochastic nonlinear filtering, Mitter [23] mentions that the invertibility of the Hessian of an analogous value function is related to observability. In his derivation of the extended Kalman filter, Mitter needs to invert this Hessian. Theorem 5.2 establishes this relationship in the deterministic case. Note that (5.8) is true for all $x_0 \in \Re^n$ and $t \geq T$ if $\Delta^u_T \equiv \{0\}$ in $\Re^n$.

6. Bilinear systems.

Consider the bilinear control system

$$\begin{aligned} &\dot{x}(t) = \left(A_0 + \sum_{i=1}^{m} u_i(t)A_i\right) x(t), \quad t > 0, \;\; x(0) = x_0; \\ &y(t) = Cx(t), \qquad t \geq 0, \end{aligned} \tag{6.1}$$

where A_i and C are appropriately sized matrices. For $u \in \mathcal{U}$ write $A^u(t) = A_0 + \sum_{i=1}^m u_i(t)A_i$. With $\mu(dx) = (2\pi)^{-n/2}\exp(-\frac{1}{2}|x|^2)dx$, the robust filter for the random perturbation of (6.1) has conditional density

(6.2)
$$\pi^{\epsilon}_{u,\mu}(x, t \mid \xi^{\epsilon}) = (2\pi \det(P^{\epsilon}(t)/\epsilon)^{-1})^{-n/2} \exp\left(-\frac{1}{2\epsilon}(x - \hat{x}^{\epsilon}(t))'P^{\epsilon}(t)(x - \hat{x}^{\epsilon}(t))\right),$$

where

$$\begin{aligned} &d\hat{x}^{\epsilon}(t) = A^u(t)\hat{x}^{\epsilon}(t)dt + (P^{\epsilon}(t))^{-1}C'(d\xi^{\epsilon}(t) - C\hat{x}^{\epsilon}(t)dt), \; t > 0, \;\; \hat{x}^{\epsilon}(0) = 0, \\ &\dot{P}^{\epsilon}(t) = -P^{\epsilon}(t)A^u(t) - A^u(t)'P^{\epsilon}(t) - P^{\epsilon}(t)^2 + C'C, \; t > 0, \;\; P^{\epsilon}(0) = \epsilon I. \end{aligned} \tag{6.3}$$

Note that $P^{\epsilon}(t) > 0$ for all $t \geq 0$. The observer equation (4.4) has the explicit solution

$$m(x, t) \;=\; \tfrac{1}{2}x'P(t)x + M(t)x + \tfrac{1}{2}N(t), \tag{6.4}$$

where

$$\begin{aligned} &\dot{P}(t) = -P(t)A^u(t) - A^u(t)'P(t) - P(t)^2 + C'C, \quad t > 0, \; P(0) = C'C; \\ &\dot{M}(t) = -M(t)\left(A^u(t) + P(t)\right) - y(t)'C, \quad t > 0, \quad M(0) = -y(0)'C; \\ &\dot{N}(t) = -M(t)M(t)' + \mid y(t) \mid^2, \quad t > 0, \; N(0) = \mid y(0) \mid^2 . \end{aligned} \tag{6.5}$$

If u is universal on $[0, T\,]$, then if $t \geq T$ we have $P(t) > 0$ and

$$\hat{x}(t) = \{-P(t)^{-1}M(t)'\} = \{x(t)\}. \tag{6.6}$$

In general, if $0 \leq t \leq T$, $P(t)$ need not be invertible, and $\hat{x}(t) = x(t) + \ker P(t)$, an affine subspace of $\Re^n$. Also, $\mathbf{I}_t^u(x_0) = x_0 + \ker \mathcal{O}^u(t)$, where $\mathcal{O}^u$ is the observability grammian of the time-varying pair (C, A^u). The ORC will be satisfied provided (Isidori [14])

$$\Omega^{\perp} = \bigcap_{k=0}^{n-1} \bigcap_{j_1,\dots,j_k=0}^{m} \ker(CA_{j_1}\dots A_{j_k}) = \{0\}, \tag{6.7}$$

in which case (6.1) will be observable and the finite-dimensional observer (6.5), (6.6) will compute the state exactly for any universal control (such controls are generic; Sussmann [27]).

7. Conclusion. We have shown that a finite time observer exists and computes the state exactly (if a universal input is used) for any control system satisfying minor technical conditions. The major disadvantage is that the observers are in general infinite-dimensional when the state space has infinitely many points. This is similar to the situation for stochastic nonlinear filtering theory. Our design is simple and general, but it is computationally difficult. From an engineering point of view, computational and practical questions need to be addressed and resolved to yield usable designs. We note that computational methods are currently under development, and they promise to be robust with respect to modeling uncertainties (James [16]). In addition, since the observer is the limit of a family of nonlinear filters, the consistency results suggest that the observer should tolerate system noise fairly well.

Acknowledgments. I wish to thank Professor J. S. Baras for providing the initial inspiration for this work and for encouraging me to work on this topic, and Professors L. C. Evans and W. H. Fleming for several helpful discussions.

REFERENCES

[1] J. S. BARAS, personal communication, 1985.

[2] J. S. BARAS, A. BENSOUSSAN, AND M. R. JAMES, *Dynamic observers as asymptotic limits of recursive filters: special cases*, SIAM J. Applied Math., 48 (1988), pp. 1147–1158.

[3] J. S. BARAS AND P. S. KRISHNAPRASAD, *Dynamic observers as asymptotic limits of recursive filters*, IEEE Proc. 21st CDC (1982), pp. 1126–1127.

[4] D. BESTLE AND M. ZEITZ, *Canonical form observer design for nonlinear time variable systems*, Internat. J. Control, 38 (1983), pp. 419–431.

[5] J.-M. BISMUT, *Large Deviations and the Malliavin Calculus*, Birkhauser, Boston, 1984.

[6] F. CELLE, J. P. GAUTHIER, D. KAZAKOS, AND G. SALLET, *Synthesis of nonlinear observers: a harmonic analysis approach*, Math. System Theory, 22 (1989), pp. 291–322.

[7] M. G. CRANDALL, L. C. EVANS, AND P. L. LIONS, *Some properties of viscosity solutions of Hamilton–Jacobi equations*, Trans. AMS, 282 (1984), pp. 487–502.

[8] M. H. A. DAVIS, *On a multiplicative functional transformation arising in nonlinear filtering theory*, Z. Wahrsch. verw. Gebiete, 54 (1980), pp. 125–129.

[9] W. H. FLEMING, *The Cauchy problem for a nonlinear first order Partial Differential Equation*, J. Differential Equations, 5 (1969), pp. 515–530.

[10] M. I. FREIDLIN AND A. D. WENTZELL, *Random Perturbations of Dynamical Systems*, Springer-Verlag, New York, 1984.

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

[12] O. HIJAB, *Minimum energy estimation*, Ph.D. dissertation, University of California, Berkeley, 1980.

[13] ———, *Asymptotic Bayesian estimation of a first order equation with a small diffusion*, Ann. Probab., 12 (1984), pp. 890–902.

[14] A. ISIDORI, *Nonlinear Control Systems: An Introduction*, Springer-Verlag, New York, 1985.

[15] M. R. JAMES, *Asymptotic nonlinear filtering and large deviations with application to observer design*, Ph.D. dissertation, University of Maryland (SRC Tech. Report Ph.D. 88-1), 1988.

[16] ———, *A numerical method for finite time observers*, preprint, 1990.

[17] M. R. JAMES AND J. S. BARAS, *Nonlinear filtering and large deviations: A* PDE-*control theoretic approach*, Stochastics, 23 (1988), pp. 391–412.

[18] D. JI, *Asymptotic analysis of nonlinear filtering problems*, Ph.D. dissertation, Brown University, Providence, RI, 1987.

[19] R. E. KALMAN, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mex., 1960, pp. 785–119.

[20] A. J. KRENER AND W. RESPONDEK, *Nonlinear observers with linearizable error dynamics*, SIAM J. Control Optim., 23 (1985), pp. 197–216.

[21] S. R. KUO, D. ELLIOTT, AND T. J. TARN, *Exponential observers for nonlinear dynamic systems*, Inform. and Control, 29 (1975), pp. 204–216.

[22] D. G. LUENBERGER, *Observers for multivariable systems*, IEEE Trans. Automat. Control, 11 (1966), pp. 190–197.

[23] S. K. MITTER, *Approximations for nonlinear filtering*, NATO Adv. Stud. Inst., Algrave, Portugal, 1980.

[24] R. E. MORTENSEN, *Maximum-likelihood recursive nonlinear filtering*, J. Opt. Theory Appl., 2 (1968), pp. 386–394.

[25] W. RUDIN, *Principles of Mathematical Analysis*, McGraw-Hill, New York, 1964.

[26] E. D. SONTAG, *On the observability of polynomial systems,* I: *finite-time problems*, SIAM J. Control Optim., 17 (1979), pp. 139–151.

[27] H. J. SUSSMANN, *Single-input observability of continuous-time systems*, Math. Systems Theory, 12 (1979), pp. 371–393.

[28] S. R. S. VARADHAN, *Asymptotic probabilities and differential equations*, Comm. Pure Appl. Math., 19 (1966), pp. 261–286.

[29] B. L. WALCOTT, M. J. CORLESS, AND S. H. ZAK, *Comparative study of nonlinear state observation techniques*, Internat. J. Control, 45 (1987), pp. 2109–2132.

[30] D. WILLIAMSON, *Observation of bilinear systems with application to biological control*, Automatica, 13 (1977), pp. 243–254.

[31] W. M. WONHAM, *Linear Multivariable Control: a Geometric Approach*, Springer-Verlag, New York, 1979.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 4, pp. 968–998, July 1991

AN EXACT PENALIZATION VIEWPOINT OF CONSTRAINED OPTIMIZATION*

JAMES V. BURKE†

Abstract. In their seminal papers Eremin [*Soviet Mathematics Doklady*, 8 (1966), pp. 459–462] and Zangwill [*Management Science*, 13 (1967), pp. 344–358] introduce a notion of exact penalization for use in the development of algorithms for constrained optimization. Since that time, exact penalty functions have continued to play a key role in the theory of mathematical programming. In the present paper, this theory is unified by showing how the Eremin–Zangwill exact penalty functions can be used to develop the foundations of the theory of constrained optimization for finite dimensions in an elementary and straightforward way. Regularity conditions, multiplier rules, second-order optimality conditions, and convex programming are all given interpretations relative to the Eremin–Zangwill exact penalty functions. In conclusion, a historical review of those results associated with the existence of an exact penalty parameter is provided.

Key words. exact penalty functions, calmness, constraint qualification, optimality conditions, convex programming

AMS(MOS) subject classifications. 49A42, 49D30, 49D37, 90D30

1. Introduction.

In their seminal papers Eremin [23] and Zangwill [75] introduced a notion of exact penalization for use in the development of algorithms for nonlinear constrained optimization. This notion of exact penalization is the natural extension of the so-called big-M method of linear programming (see Charnes, Cooper, and Henderson [14, §4] for the earliest reference known to us) to nonlinear programming. Since that time, exact penalty functions have continued to play a key role in the theory of mathematical programming. Within the algorithmic sphere, the history of these functions is quite rich, even though their use has been, and still is, a topic of controversy. The root of this controversy is the nondifferentiable nature of these functions. From an algorithmic viewpoint, this nondifferentiability can induce the so-called Maratos effect (a phenomenon that prevents rapid local convergence). A great deal of effort has been devoted to overcoming this difficulty, leading to the development of the so-called watchdog technique [12] and second-order correction techniques [19], [28], [26], [29], and others. Other authors, in an effort to avoid the problems associated with nondifferentiability, have introduced entirely different classes of exact penalty functions that are differentiable [5], [30], [34], [60], and [69]. The research in this area continues at a rapid pace and the controversies over the use of nondifferentiable exact penalty functions in algorithms are far from nearing resolution. This paper can, in many ways, be viewed as a contribution to this discussion. However, our approach is from a rather different perspective. We do not discuss algorithms at all, rather we demonstrate how the Eremin–Zangwill exact penalty functions can be used to develop the foundations of the theory of constrained optimization in an elementary and straightforward way. In doing so, we show how all of the fundamental notions and results in constrained optimization can be derived from the Eremin–Zangwill exact penalty functions, from regularity conditions such as calmness [15], [66], to the

*Received by the editors September 16, 1986; accepted for publication (in revised form) June 1, 1990.

†Mathematics Department, GN-50, University of Washington, Seattle, Washington 98195. The work of the author was supported in part by National Science Foundation grant DMS–8803206 and by Air Force Office of Scientific Research grant AFOSR–860080.

existence of Lagrange multipliers, to second-order necessary and sufficient conditions for optimality. The derivation of these results by means of the Eremin–Zangwill exact penalty functions is by no means strained or artificial, quite the contrary, the proofs are often simplified at the expense of obtaining a more powerful result. Thus, our goal in this endeavor is not to demonstrate the viability of these penalty functions for use in algorithmic development, but rather to demonstrate their role vis-á-vis the foundations of the theory and to provide an interpretation for many of the familiar objects in this theory in terms of the corresponding objects associated with these penalty functions. Hopefully, one consequence of these investigations is that the practical significance of these penalty functions can be more accurately assessed.

We begin §2 by reviewing some of the fundamental results and concepts associated with constrained optimization. We discuss calmness, regularity, constraint qualifications, and their relationships vis-á-vis exact penalization. This section contains all of the first-order results related to the existence of Kuhn–Tucker [43] multipliers. In §3 we show how exact penalization techniques can be used to derive a multiplier theorem in the absence of a constraint qualification. This multiplier rule is reminiscent of the one given by John [42]. Second-order results are obtained in §4. The case of convex programming is studied in §5, and in §6 we provide a historical review of the literature on the existence of a finite exact penalty parameter. The approach to the theory of constrained optimization from the viewpoint of exact penalization is also the theme of Fletcher [29, §14.3], Garcia-Palomares [31], and Rockafellar [64]. A very nice survey of exact penalization techniques in general is given by Fletcher [27]. The present paper is based on Burke [9], wherein several further results and generalizations are obtained.

The notation that we employ is for the most part standard; however, a partial list is provided for the reader's convenience. Let X be a real normal linear space and let X^* be its topological dual. The spaces X and X^* are paired in duality by the continuous bilinear form

$$\langle x^*, x\rangle := x^*(x)$$

defined on $X^* \times X$. Given $x_1, x_2 \in X$ the line segment joining them is denoted by

$$[x_1, x_2] := \{\lambda x_1 + (1-\lambda)x_2 : \lambda \in [0,1]\}.$$

Let C be a subset of X. Then cl(C) is the closure of C, int(C) is the interior of C, and ri(C) is the interior of C relative to its affine hull, i.e., the smallest closed affine set containing C. The core of C, denoted core(C), is the set of all point $z \in C$ such that every line through z contains a line segment $[z_1, z_2]$ with $z \in [z_1, z_2] \subset C$ and $z_1 \neq z \neq z_2$. In finite dimensions, we have core$(C) =$ int(C). The *polar* of C is given by

$$C^0 := \{x^* \in X^* : \langle x^*, x\rangle \leq 1 \text{ for all } x \in C\}$$

and the *positive conjugate* of C is $C^* := -C^0$. The *recession cone* of C is

$$\text{rec}(C) := \{y \in X : C + y \subset cl(C)\}$$

and the cone generated by C is

$$\text{cone}(C) := \cup_{\lambda \geq 0} \lambda C,$$

where for any two subsets S_1 and S_2 of X and any two scalars $\alpha, \beta \in \mathbb{R}$ we have

$$\alpha S_1 + \beta S_2 := \{\alpha s_1 + \beta s_2 : s_1 \in S_1, s_2 \in S_2\}.$$

The *support* and *indicator* functions for C are given, respectively, as

$$\psi^*(x^*|C) := \sup\{\langle x^*, y\rangle : y \in C\}$$

and

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

The *barrier cone* of C is

$$\text{bar}(C) := \{x^* \in X^* : \psi^*(x^*|C) < \infty\}$$

and the relation

$$\text{rec}(C) = [\text{bar}(C)]^0$$

holds if X is reflexive.

A *multifunction* T mapping X into Y where Y is another real normal linear space, written $T : X \rightrightarrows Y$, is a mapping of X whose values are subsets of Y. The *domain* of T is the set $\text{dom}(T) := \{x \in X : T(x) \neq \emptyset\}$ and the *graph* of T is graph $(T) := \{(x, y) | y \in T(x)\}$. T is said to be *upper semicontinuous* if graph (T) is closed in $X \times Y$ under the product topology. The space $\mathcal{L}(X, Y)$ is the space of continuous linear maps from X to Y. Given $T \in \mathcal{L}(X, Y)$ we write

$$\text{ran}(T) := \{y \in Y : \exists x \in X \text{ with } y = Tx\}$$

and

$$\ker(T) := \{x \in X : Tx = 0\}.$$

If X and Y are finite-dimensional, then, with respect to fixed bases for X and Y, one can identify $\mathcal{L}(X, Y)$ with $\mathbb{R}^{m \times n}$ the set of $m \times n$ matrices, where $\dim(X) = n$ and $\dim(Y) = m$. The *adjoint* of $A \in \mathcal{L}(X, Y)$ is the uniquely defined mapping $A^* \in \mathcal{L}(Y^*, X^*)$ for which

$$\langle A^* y^*, x\rangle = \langle y^*, Ax\rangle$$

for all $(y^*, x) \in Y^* \times X$. In finite dimensions we have $A^* = A^T$.

Let $f : X \to \overline{\mathbb{R}}$ where $\overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$, we write

$$\begin{aligned} \text{dom}(f) &:= \{x \in X : f(x) < +\infty\}, \\ \text{lev}_f(x) &:= \{y \in X : f(y) \leq f(x)\}, \text{ and} \\ \text{epi}(f) &:= \{(\mu, x) : f(x) \leq \mu\}. \end{aligned}$$

We say that f is *lower semicontinuous* if $\text{epi}(f)$ is a closed set. If f is Lipschitz near a point $x \in X$, then the *Clarke generalized directional derivative*,

$$f^0(x; d) := \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y + tv) - f(y)}{t},$$

exists at x for every $d \in X$.

The norm on X is denoted $\|\cdot\|$ and its unit ball is $\mathbb{B} := \{x : \|x\| \leq 1\}$. The dual norm is given by $\|x^*\|_0 := \psi^*(x^*|\mathbb{B})$ and its unit ball is $\mathbb{B}^0$. The distance function for a set $C \subset X$ is given by

$$\text{dist}(y|C) := \inf\{\|y - x\| : x \in C\}.$$

For $C \subset X^*$, the dual distance function is denoted

$$\mathrm{dist}_0(y|C) := \inf\{\|y - x\|_0 : x \in C\}.$$

In finite dimensions the Euclidean norm plays a special role and is denoted by $\|\cdot\|_2$ with corresponding distance function $\mathrm{dist}_2(\cdot|C)$. The distance function for a set $C \subset X$ is Lipschitz with Lipschitz constant 1, and so its Clarke generalized directional derivative exists at every point in all directions. Based on this observation we define the *tangent cone* to a point $x \in C$ by

$$T(x|C) := \{d \in X : \mathrm{dist}(\cdot|C)^0(x; d) = 0\}$$

with the *normal cone* defined via polarity

$$N(x|C) := T(x|C)^0.$$

For convex sets, these objects reduce to the usual notions of tangent and normal cone. In finite dimensions one can also define the *limiting proximal normal cone* at a point $x \in C$ by $\widehat{N}(x|C) := \{\lambda \lim v_i/\|v_i\| : \lambda \geq 0, v_i \perp C \text{ at } x_i \to x,\ v_i \to 0\}$, where one writes $v \perp C$ at y to mean that $y \in \mathrm{cl}(C)$ and $v = y^1 - y$ with $\|y^1 - y\|_2 = \mathrm{dist}_2(y|C)$. One has that $N(x|C)$ is the closed convex hull of $\widehat{N}(x|C)$.

Given $f : X \to \overline{\mathbb{R}}$ the *generalized subdifferential* of f at $x \in \mathrm{dom}(f)$ is given by

$$\partial f(x) := \{x^* \in X^* : (-1, x^*) \in N((f(x), x)|\mathrm{epi}(f))\},$$

the *asymptotic subdifferential* is

$$\partial^\infty f(x) := \{x^* \in X^* : (0, x^*) \in N((f(x), x)|\mathrm{epi}(f))\},$$

the *limiting proximal subdifferential* is

$$\widehat{\partial} f(x) := \{x^* \in X^* : (-1, x^*) \in \widehat{N}((f(x), x)|\mathrm{epi}(f))\},$$

and the *asymptotic limiting proximal subdifferential* is

$$\widehat{\partial}^\infty f(x) := \{x^* \in X : (0, x^*) \in \widehat{N}((f(x), x)|\mathrm{epi}(f))\}.$$

Clearly, $\partial^\infty f(x) = \mathrm{rec}(\partial f(x))$ whenever $\partial f(x) \neq \emptyset$. The *generalized directional derivative* of f is then defined to be

$$f^0(x; v) := \psi^*(v|\partial f(x))$$

with $f^0(x; v) := -\infty$ if $\partial f(x) = \emptyset$. This notation is consistent with that of the Clarke subdifferential for locally Lipschitz functions.

The function f is said to be *subdifferentially regular* at a point $x \in \mathrm{dom}(f)$ if

$$\liminf_{\substack{u \to v \\ t \downarrow 0}} \frac{f(x + tu) - f(x)}{t} = f^0(x; v)$$

for all $v \in X$, in which case

$$f^0(x; v) = f'(x; v) := \lim_{t \downarrow 0} \frac{f(x + tv) - f(x)}{t}.$$

A function $F : X \to Y$ has *Frechet derivative* $F'(x) \in \mathcal{L}(X,Y)$ at $x \in X$ if

$$F(y) = F(x) + F'(x)(y - x) + o(\|y - x\|),$$

where $\lim_{y \to x} o(\|y - x\|)/\|y - x\| = 0$. The mapping F is *strictly differentiable* at $x \in X$ if there exists $F_s'(x) \subset \mathcal{L}(X,Y)$ such that

$$\lim_{\substack{x' \to x \\ t \downarrow 0}} \frac{F(x' + tv) - F(x')}{t} = F_s'(x)v$$

for all v in X. If $f : X \to \overline{R}$ is strictly differentiable at a point $x \in \mathrm{dom}(f)$, then $\partial f(x) = \{f_s'(x)\}$.

If both X and Y are finite-dimensional and $F : X \to Y$ is locally Lipschitz, then F is almost everywhere differentiable in the sense of Lebesgue measure. The *generalized Jacobian* of F at a point $x \in X$, denoted $\partial F(x)$, is the convex hull of all operators in $\mathcal{L}(X,Y)$ obtained as the limit of sequences of the form $\{F'(x_i)\}$ where $x_i \to x$ and $F'(x_i)$ exists at each x_i. Again, if F is strictly differentiable at x, then $\partial F(x) := \{F_s'(x)\}$.

Let $f : X \to \overline{\mathbb{R}}$ and $C \subset X$. We write

$$\arg\min\{f(x) : x \in C\} := \{x \in C : f(x) = \min\{f(x) : x \in C\}\}$$

and define $\arg\max\{f : x \in C\}$ similarly. A local minimum of *radius* ε for the problem $\min\{f(x) : x \in C\}$ is any point $x \in C$ such that $f(x) \leq f(y)$ for all $y \in C \cap (x + \varepsilon\mathbb{B})$.

For more information about the objects defined above see [15]–[17], [54], and [65]–[68].

2. The fundamentals: calmness, regularity, and exact penalization. Let X and Y be normal linear spaces and consider the problem

$$(\mathcal{P}) \qquad \begin{array}{l} \text{minimize } f(x) \\ \text{subject to } g(x) \in C, \end{array}$$

where $f : X \to \overline{\mathbb{R}} := \mathbb{R} \cup \{+\infty\}$, $g : X \to Y$, and C is a closed subset of Y. We begin with a discussion of regularity conditions that allow the development of general multiplier rules for $\mathcal{P}$. One of the weakest such conditions was proposed by Rockafellar and is known as calmness.

DEFINITION 2.1. Let f, g, X, Y, and C be as in the statement of $\mathcal{P}$ and consider the perturbed problems

$$(\mathcal{P}_u) \qquad \begin{array}{l} \text{minimize } f(x) \\ \text{subject to } g(x) \in C + u. \end{array}$$

Let $\overline{x} \in X$ and $\overline{u} \in Y$ be such that $g(\overline{x}) \in C + \overline{u}$ and $\overline{x} \in \mathrm{dom}(f) := \{x \in X : f(x) < +\infty\}$. The problem $\mathcal{P}_{\overline{u}}$ is said to be *calm* at $\overline{x}$ if there are constants $\overline{\alpha} \geq 0$ and $\varepsilon > 0$ such that for every pair $(x, u) \in X \times Y$ with $\|x - \overline{x}\| \leq \varepsilon$ and $g(x) \in C + u$ we have

$$f(x) + \overline{\alpha}\|u - \overline{u}\| \geq f(\overline{x}). \tag{2.1}$$

The constants $\overline{\alpha}$ and ε are called the *modulus* and *radius* of calmness for $\mathcal{P}_{\overline{u}}$ at $\overline{x}$, respectively.

The family of perturbed problems $\mathcal{P}_u$ is said to be calm at $\overline{u}$ if

$$\liminf_{u \to \overline{u}} \frac{V(u) - V(\overline{u})}{\|u - \overline{u}\|} > -\infty, \tag{2.2}$$

where

$$V(u) := \begin{cases} +\infty, & \text{if } \{x : g(x) \in C + u\} = \emptyset \\ \min\{f(x) : g(x) \in C + u\}, & \text{otherwise} \end{cases}$$

is the value function for the family $\mathcal{P}_u$.

Remarks. (1) This definition for $\mathcal{P}_{\overline{u}}$ to be calm at $\overline{x}$ varies from the definition that is usually given (eg., see Clarke [15, Def. 6.4.1]); however, in Burke [8, §2], it is shown that they are equivalent when g is continuous at $\overline{x}$.

(2) Observe that if $\mathcal{P}_{\overline{u}}$ is calm at $\overline{x}$, then $\overline{x}$ is necessarily a local solution to $\mathcal{P}_{\overline{u}}$, and if $\mathcal{P}_u$ is calm at $\overline{u}$, then for any solution $\overline{x}$ to $\mathcal{P}_{\overline{u}}, \mathcal{P}_{\overline{u}}$ is calm at $\overline{x}$.

(3) The notion of calmness is closely related to the notion of a Φ_1-subdifferential introduced in Dolecki and Rolewicz [20].

The calmness hypothesis is quite weak and in many situations is easily verified. In finite dimensions, calmness holds on a dense subset of the perturbations.

PROPOSITION 2.1. (1) (Clarke [15, Prop. 6.4.5]) *Suppose that* $Y := \mathbb{R}^m, C := \mathbb{R}^m_-$, *and* $f := f_0 + \psi(\cdot|S)$ *with* $S \subset X$ *nonempty and closed, and* $f_0 : X \to \mathbb{R}$ *and* $g : X \to \mathbb{R}^m$ *locally Lipschitzian. If* $V(u)$ *is finite for all* u *near* 0, *then for almost all* u *in a neighborhood of the origin the problem* $\mathcal{P}_u$ *is calm.*

(2) (Burke [8, Prop. 3.1]) *Suppose that* Y *is finite-dimensional,* f *is lower semicontinuous, and* g *is continuous. If* $\overline{u} \in Y$ *and* $\gamma > 0$ *are such that* V *is bounded on* $\overline{u} + \gamma\mathbb{B}$, *then* $\mathcal{P}_u$ *is calm on a dense subset of* $\overline{u} + \gamma\mathbb{B}$.

From (2.2) it is clear that calmness is a weak variational property of the value function V. A condition of this type is always required for establishing the existence of multipliers. It is remarkable that the notion of calmness at a solution to $\mathcal{P}_{\overline{u}}$ is equivalent to the existence of a finite exact penalty parameter.

THEOREM 2.1 (Burke [8, Thm. 1.1]). *Let* $\overline{x} \in X$ *and* $\overline{u} \in Y$ *be such that*

$$g(\overline{x}) \in C + \overline{u} \quad \text{and} \quad \overline{x} \in \operatorname{dom}(f).$$

Then $\mathcal{P}_{\overline{u}}$ *is calm at* $\overline{x}$ *with modulus* $\overline{\alpha} \geq 0$ *and radius* $\varepsilon > 0$ *if and only if* $\overline{x}$ *is a local minimum of radius* ε *for*

$$P_{\overline{u},\alpha}(x) := f(x) + \alpha \operatorname{dist}(g(x)|C + \overline{u})$$

for all $\alpha \geq \overline{\alpha}$, *that is,*

$$P_{\overline{u},\alpha}(\overline{x}) \leq P_{\overline{u},\alpha}(x)$$

for all $x \in \overline{x} + \varepsilon\mathbb{B}$ *and* $\alpha \geq \overline{\alpha}$.

Remark. The fact that calmness implies the existence of an exact penalty parameter is also established in Clarke [15] and Dolecki and Rolewicz [20]. However, the reverse implication and the precision of this correspondence is first established in [8].

Thus, at this early juncture we see that the Eremin–Zangwill exact penalty functions play a fundamental role in the theory. Under the calmness hypothesis we can obtain multiplier rules for $\mathcal{P}$ by first invoking Theorem 2.1 and then applying the pertinent calculus rules of an appropriate subdifferential (e.g., the Clarke subdifferential [15]–[17], the Michel–Penot subdifferential [53], the limiting proximal subdifferential [65]–[66], etc.)

We present two sample results based on the subdifferential calculus developed in Clarke [15] and Rockafellar [66], [68].

THEOREM 2.2. (1) *Suppose* $\mathcal{P}$ *is calm at* $\overline{x} \in X$, g *is strictly differentiable at* $\overline{x}$ *with strict derivative* $g_s'(\overline{x})$, *and* $\partial f(\overline{x}) \neq \emptyset$. *Then there is a* $y \in N(g(\overline{x})|C)$ *such that*

$$0 \in \partial f(\overline{x}) + g_s'(\overline{x})^* y.$$

(2) *If* X *and* Y *are finite-dimensional,* $\mathcal{P}$ *is calm at* $\overline{x} \in X$, f *is lower semicontinuous near* $\overline{x}$, *and* g *is Lipschitzian near* $\overline{x}$, *then there exists* $y \in N(g(\overline{x})|C)$ *such that*

$$0 \in \partial f(\overline{x}) + \partial g(\overline{x})^* y.$$

Proof. (1) By Theorem 2.1, $\overline{x}$ is a local minimum for $P_\alpha(x) := f(x) + \alpha \operatorname{dist}(g(x)|C)$ for all α sufficiently large. Hence $0 \in \partial P_\alpha(\overline{x})$ for all $\alpha \geq \overline{\alpha}$ for some $\overline{\alpha} \geq 0$. By [68, Cor. 2],

$$\partial P_\alpha(\overline{x}) \subset \partial f(\overline{x}) + \alpha\partial\,[\operatorname{dist}(g(\cdot)|C)](\overline{x}).$$

From [15, Prop. 2.4.2],

$$N(g(\overline{x})|C) = cl[\cup_{\lambda \geq 0} \lambda\partial\,[\operatorname{dist}(\cdot|C)](g(\overline{x}))]. \tag{2.3}$$

Consequently, by the chain rule [15, Thm. 2.3.10],

$$0 \in \partial f(\overline{x}) + g_s'(\overline{x})^* N(g(\overline{x})|C),$$

from which the result follows.

(2) This is an immediate consequence of Rockafellar [66, Cor. 5.2.3] and inclusion (2.3). □

Remarks. (1) To incorporate an abstract constraint of the form $x \in S \subset X$ we simply replace f by $f + \psi(\cdot|S)$.

(2) We do not claim that the results in Theorem 2.2 are original. Results similar to these can be found elsewhere in the literature, e.g., [1], [3], [15]–[17], [39]–[41], [45], [54], [63]–[66]. However, the proofs that we provide are different from those that are usually provided, due to the explicit dependence on Theorem 2.1.

Various conditions can be found in the literature that ensure that the calmness hypothesis is satisfied. All of these conditions are related to the regularity of the constraint systems of the form

$$g(x) \in C \quad \text{and} \quad x \in S \subset X. \tag{2.4}$$

DEFINITION 2.2. System (2.4) is said to be *regular* at a solution x_0 if there exist constants $\kappa > 0$ and $\varepsilon > 0$ such that

$$\operatorname{dist}(x|\Omega(u)) \leq \kappa \operatorname{dist}(g(x)|C + u)$$

for all $x \in (x_0 + \varepsilon\mathbb{B}) \cap S$ and $u \in \varepsilon\mathbb{B}$ where

$$\Omega(u) := \{x \in X : g(x) \in C + u, x \in S\}.$$

The constant κ is called the *modulus of regularity* for (2.4) at x_0.

Remark. This and more general notions of regularity for (2.4) are studied by several authors, e.g., [1], [4], [7], [15], [20], [48], [49], [51], [52], [61], [62], [66]–[68], [73].

Calmness and regularity are related via Clarke's elementary exact penalization theorem.

THEOREM 2.3 (Clarke [15, Prop. 2.4.3]). *Let $f : X \to \overline{\mathbb{R}}$ be Lipschitz of rank κ on a set $T \subset X$. Let $\overline{x}$ belong to a set $\Omega \subset T$ and suppose that f attains a minimum over Ω at $\overline{x}$. Then for any $\widehat{\kappa} \geq \kappa$, the function $\pi_{\widehat{\kappa}}(x) := f(x) + \widehat{\kappa}\ \mathrm{dist}(x|\Omega)$ attains a minimum over T at $\overline{x}$. If $\widehat{\kappa} > \kappa$ and Ω is closed, then any other point minimizing $\pi_{\widehat{\kappa}}$ over T must also lie in Ω.*

We have the following elementary corollaries to Theorem 2.3.

COROLLARY 2.3.1. *Consider the problem $\mathcal{P}$ with g continuous and $f := f_0 + \psi(\cdot|S)$ for some $f_0 : X \to \mathbb{R}$ and some $S \subset X$ closed and nonempty. Suppose that $\overline{x} \in S$ is a local solution to $\mathcal{P}$ at which the system (2.4) is regular with modulus κ_1 and near which f_0 is Lipschitz of rank κ_2, then $\overline{x}$ is a local minimum of $P_\alpha(x) := f(x) + \alpha\ \mathrm{dist}(g(x)|C)$ for all $\alpha \geq \kappa_1\kappa_2$. If $\alpha > \kappa_1\kappa_2$, then there is a neighborhood of $\overline{x}$ such that any other local minimum $\widehat{x}$ of $P_\alpha(x)$ within this neighborhood is such that $f(\overline{x}) = f(\widehat{x})$ and $g(\widehat{x}) \in C$.*

Proof. Let $\varepsilon > 0$ be such that f_0 is Lipschitz of rank κ_2 on $\overline{x} + \varepsilon\mathbb{B}$, the defining inequality for regularity holds for all $x \in (\overline{x} + \varepsilon\mathbb{B}) \cap S$ and $u \in \varepsilon\mathbb{B}$, and $f(\overline{x}) \leq f(x)$ for all $x \in \{z : g(z) \in C\} \cap (\overline{x} + \varepsilon\mathbb{B})$. Set $\Omega := \Omega(0) \cap (\overline{x} + \varepsilon\mathbb{B})$ and note that since g is continuous, the set Ω is closed. By Theorem 2.3, $\pi_{\widehat{\kappa}}(x)$ attains a minimum over $\overline{x} + \varepsilon\mathbb{B}$ at $\overline{x}$ for $\widehat{\kappa} \geq \kappa_2$, and if $\widehat{\kappa} > \kappa_2$, then any other minimum of $\pi_{\widehat{\kappa}}$ over $\overline{x} + \varepsilon\mathbb{B}$ must also lie in Ω. Then, for every $\delta \in (0, \frac{1}{3}\varepsilon)$ and $y \in \overline{x} + \frac{1}{3}\varepsilon\mathbb{B}$, there is a $z_\delta \in \Omega(0)$ such that $\|y - z_\delta\| \leq \mathrm{dist}(y|\Omega(0)) + \delta \leq \frac{2}{3}\varepsilon$. Hence $\|z_\delta - \overline{x}\| \leq \|y - z_\delta\| + \|y - \overline{x}\| \leq \varepsilon$ so that $\mathrm{dist}(y|\Omega) \leq \mathrm{dist}(y|\Omega(0)) + \delta$. Letting $\delta \downarrow 0$ we find that $\mathrm{dist}(y|\Omega) = \mathrm{dist}(y|\Omega(0))$ for all $y \in \overline{x} + \frac{1}{3}\varepsilon\mathbb{B}$. The result now follows from the definition of regularity with the neighborhood of $\overline{x}$ being $\overline{x} + \frac{1}{3}\varepsilon\mathbb{B}$. □

COROLLARY 2.3.2. *Consider the problem $\mathcal{P}$ and let f, g, and $\overline{x}$ be as in Corollary 2.3.1. Then $\mathcal{P}$ is calm at $\overline{x}$.*

Proof. This is an immediate consequence of Corollary 2.3.1 and Theorem 2.1. □

Remark. Dolecki and Rolewicz [20] obtain a result similar to Corollary 2.3.1 in a more general setting by using somewhat different techniques. Their result is based upon the notion of an upper Hausdorff semicontinuous multifunction.

Conditions yielding the regularity of the constraint system (2.4) have been studied by many authors [1], [4], [7], [15], [20], [48], [49], [51], [52], [61], [62], [66]–[68], [73]. The first and most famous of these results is the Lyusternik theorem [48]. An excellent discussion of a variety of these regularity results is given in Borwein [7]. In the mathematical programming literature such conditions are often called constraint qualifications, e.g., the Mangasarian–Fromovitz constraint qualification [51], [52]. In his thesis, Maguregui [49, Chap. 2], introduced the constraint qualification

$$0 \in \mathrm{core}(g(x_0) + g'(x_0)(S - x_0) - C). \tag{2.5}$$

THEOREM 2.4 (Maguregui [49, Chap. 2]). *Suppose that X and Y are Banach spaces, the sets $S \subset X$ and $C \subset Y$ are nonempty, closed, and convex, and $g : X \to Y$ is strictly differentiable at $\overline{x} \in S$. If $g(\overline{x}) \in C$ and (2.5) is satisfied, then system (2.4) is regular at $\overline{x}$.*

Remarks. (1) Using the constraint qualification

$$g_s'(x_0)T(x_0|S) - T(g(x_0)|C) = Y, \tag{2.6}$$

Borwein [7, Thm. 4.3] show that the convexity assumption on the sets C and S can be removed if we instead assume that the sets S and C are epi-Lipschitzian (in the sense of Rockafellar [68]) at x_0 and $g(x_0)$, respectively.

(2) If X and Y are finite-dimensional or if C and S are convex, then the conditions (2.5), (2.6), and

$$\ker([g_s'(x_0)^T, I]) \cap [N(g(x_0)|C) \times N(x_0|S)] = \{0\} \tag{2.7}$$

are all equivalent. Moreover, if $S = \mathbb{R}^n$, and $C := \mathbb{R}^s \times \{0\}_{R^{m-s}}$, all of the conditions (2.5)–(2.7) are equivalent to the Mangasarian–Fromovitz constraint qualification.

(3) In the finite-dimensional case, Borwein [7, Thm. 3.2] has shown that we can generalize (2.7) to

$$\ker([\partial g(x_0)^T, I]) \cap [N(g(x_0)|C) \times N(x_0|S)] = \{0\}, \tag{2.8}$$

where

$$\ker([\partial g(x_0)^T, I]) := \{(y, z) \in \mathbb{R}^m \times \mathbb{R}^n : 0 \in \partial g(x_0)^T y + z\},$$

and still guarantee the regularity of system (2.4).

COROLLARY 2.4.1. *Let the hypotheses of Theorem* 2.4 *hold and consider the problem $\mathcal{P}$ with g continuous and $f := f_0 + \psi(\cdot|S)$, where $f_0 : X \to \mathbb{R}$ Lipschitz near $\overline{x}$. Then $\mathcal{P}$ is calm at $\overline{x}$, or equivalently, $\overline{x}$ is a local minimum for P_α for all α sufficiently large. Moreover, there is a threshold value of α, say $\overline{\alpha}$, and a neighborhood U of $\overline{x}$ such that if $\alpha > \overline{\alpha}$, then any other local minimum of P_α, $\widehat{x} \in U$, must satisfy $f(\overline{x}) = f(\widehat{x})$ and $g(\widehat{x}) \in C$.*

Proof. This is an immediate consequence of Theorem 2.1, Corollary 2.3.1, and Theorem 2.4. □

Remarks. (1) Corollary 2.4.1 extends Han and Mangasarian [33, Thm. 4.4] where it is assumed that $X = \mathbb{R}^n, Y = \mathbb{R}^m$, $S = \mathbb{R}^n$, and $Y := \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$, f_0 and g are continuously differentiable and $\overline{x}$ is a strict local solution to $\mathcal{P}$.

(2) Dolecki and Rolewicz [20, Thm. 2.1] obtain a result similar to Corollary 2.4.1 in a somewhat more general setting. Their result is based upon the notion of locally controllable image nearly inner approximations (inia).

In finite dimensions it is possible to strengthen the result in Corollary 2.4.1 by dropping the requirement that S be convex. Clarke establishes this in [15, Cor. 5, p. 244]. It can also be established by methods that place exact penalty techniques within a broader context of convex composite optimization. In convex composite optimization one studies the problem

$$(\mathcal{Q}) \qquad \text{minimize} \quad q(x)$$

with $q := f + h \circ g$ where $f : X \to \overline{\mathbb{R}}$ and $g : X \to Y$ are as in the statement of $\mathcal{P}$, and $h : Y \to \overline{\mathbb{R}}$ is convex. If C is convex, then $P_\alpha(x) := f(x) + \alpha \operatorname{dist}(g(x)|C)$ is an example of a convex composite function. The following result concerning $\mathcal{Q}$ is a modest extension of a result originally due to Burke and Poliquin [11, Thm. 3.1].

THEOREM 2.5. *Consider the problem $\mathcal{Q}$ where $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is lower semicontinuous, $g : \mathbb{R}^n \to \mathbb{R}^m$ is locally Lipschitz, and $h : \mathbb{R}^m \to \overline{\mathbb{R}}$ is lower semicontinuous and convex.*

Let $\overline{x} \in \operatorname{dom}(q)$ and suppose that

$$\begin{bmatrix} 0 \in \partial^\infty f(\overline{x}) + \partial g(\overline{x})^* y \\ y \in N(g(\overline{x})|\operatorname{dom}(h)) \end{bmatrix} \Longrightarrow y = 0. \tag{2.9}$$

Define

$$q_\alpha(x) := f(x) + h_\alpha(g(x)) \tag{2.10}$$

with

$$h_\alpha(y) := \inf\{h(z) + \alpha\|y - z\| : z \in \mathbb{R}^m\}. \tag{2.11}$$

If $\overline{x} \in \mathrm{dom}(q)$ *is a local solution to* $\mathcal{Q}$, *then there is an* $\overline{\alpha} > 0$ *such that* $\overline{x}$ *is a local minimizer for* $q_\alpha(x)$ *for all* $\alpha \geq \overline{\alpha}$.

Remarks. (1) The proof of Theorem 2.5 is rather technical, and so is relegated to Appendix A.

(2) The operation employed in (2.11) is known as the infimal convolution of h and $\alpha\|\cdot\|$, and is written $h \,\square\, \alpha\|\cdot\|$. In general, we have

$$\mathrm{epi}[h_1 \,\square\, h_2] = \mathrm{epi}(h_1) + \mathrm{epi}(h_2)$$

for any two convex functions h_1 and h_2. Consequently, $h_1 \,\square\, h_2$ is always convex.

(3) Note that $\mathrm{dom}(h_\alpha) = \mathbb{R}^m$ even if $\mathrm{dom}(h) \neq \mathbb{R}^m$. Hence $\mathrm{dom}(q_\alpha) = \mathrm{dom}(f)$.

If the set C in problem $\mathcal{P}$ is convex, then $\mathcal{P}$ can be seen as an instance of $\mathcal{Q}$ by taking $h := \psi(\cdot|C)$. In this case we have

$$\begin{aligned} h_\alpha(y) :&= \inf\{\psi(z|C) + \alpha\|y - z\| : z \in \mathbb{R}^m\} \\ &= \alpha \ \mathrm{dist}(y|C), \end{aligned}$$

and so

$$q_\alpha(x) = P_\alpha(x) := f(x) + \alpha \ \mathrm{dist}(g(x)|C).$$

Thus Theorem 2.5 can be used to provide conditions under which a finite exact penalty parameter α exists. Condition (2.9) is just another constraint qualification. In particular, if $f := f_0 + \psi(\cdot|S)$ with f_0 locally Lipschitz and S closed, then (2.8) and (2.9) are equivalent and we recover Clarke's result [15, Cor. 5, p. 244] as a special case. Constraint qualifications of the type (2.9) were originally formulated by Rockafellar in [66] and [68]. These comments yield the following corollary to Theorem 2.5.

COROLLARY 2.5.1. *Let* $f : \mathbb{R}^n \to \overline{\mathbb{R}}$, $g : \mathbb{R}^n \to \mathbb{R}^m$, *and* $\overline{x} \in \mathbb{R}^n$ *be as in the statement of Theorem* 2.5 *and consider problem* $\mathcal{P}$. *If* (2.9) *holds with* $h := \psi(\cdot|C)$ *where is* C *nonempty closed and convex, then* $\mathcal{P}$ *is calm at* $\overline{x}$, *or equivalently,* $\overline{x}$ *is a local minimum for* P_α *for all* α *sufficiently large. Moreover, there is a threshold value of* α, *say* $\overline{\alpha}$, *and a neighborhood* U *of* $\overline{x}$ *such that if* $\alpha > \overline{\alpha}$, *then any minimum* $\widehat{x}$ *of* P_α *on* U *must satisfy* $f(\overline{x}) = f(\widehat{x})$ *and* $g(\widehat{x}) \in C$.

Proof. In light of the comments preceding the statement of the result, we need only prove the last part of the result. To this end let $\overline{\alpha}$ be any value of α for which $\overline{x}$ is a local minimum of P_α, and let U be any neighborhood of $\overline{x}$ such that $P_{\overline{\alpha}}(\overline{x}) \leq P_{\overline{\alpha}}(x)$ for all $x \in U$. If $\widehat{\alpha} > \overline{\alpha}$, then $P_{\overline{\alpha}}(\overline{x}) \leq P_{\widehat{\alpha}}(x)$ for all $x \in U$. If $\widehat{x}$ is any other minimum of $P_{\widehat{\alpha}}$ on U, then

$$P_{\widehat{\alpha}}(\widehat{x}) = f(\overline{x}) \leq f(\widehat{x}) + \frac{\overline{\alpha} + \widehat{\alpha}}{2} \ \mathrm{dist}(g(\widehat{x})|C).$$

Consequently, $\mathrm{dist}(g(\widehat{x})|C) = 0$ and $f(\overline{x}) = f(\widehat{x})$. □

Remark. For the case in which $C := \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$ and both f and g are continuously differentiable, this result was first obtained by Han and Mangasarian in [33, Thm. 4.4]. Rosenberg [71, Prop. 1] later generalized Han and Mangasarian's result to the case in which $C := \mathbb{R}^s_- \times \{0\}_{R^{m-s}}$, $f := f_0 + \psi(\cdot|S)$, where f_0 and g are locally Lipschitz and S is nonempty and closed.

In this section we have obtained multiplier rules for $\mathcal{P}$ via the exact penalty function P_α and the calmness hypothesis. We call these multipliers Kuhn–Tucker multipliers. Given $x \in \Omega := \{x \in \mathrm{dom}(f) : g(x) \in C\}$, we denote these multipliers by

$$\text{K-T}(x) := \{y \in N(g(x)|C) : 0 \in \partial f(x) + \partial g(x)^* y\}, \tag{2.12}$$

where $\partial g(x)$ is always taken to be $g_s'(x)$ in the infinite-dimensional setting. This set is always closed and may be empty. It should be noted that this is an extension of the usual theory of Kuhn–Tucker multipliers; that is, if f and g are continuously differential and $C := \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$, then K-T (x) consists precisely of the usual Kuhn–Tucker multipliers for $\mathcal{P}$ at x [43], [51], [70].

PROPOSITION 2.6. *Suppose X and Y are normed linear spaces, C is a closed subset of Y, $f : X \to \overline{\mathbb{R}}$, and $g : X \to Y$. Let $\bar x \in \mathrm{dom}(P_\alpha)$ be such that $g(\bar x) \in C$.*

(1) *If g is strictly differentiable at $\bar x$ and $0 \in \partial P_\alpha(\bar x)$ for some $\alpha \geq 0$, then K-T $(\bar x) \neq \emptyset$.*

(2) *If X and Y are finite-dimensional, g is Lipschitz near $\bar x$, and $0 \in \partial P_\alpha(x)$ for some $\alpha \geq 0$, then K-T $(\bar x) \neq \emptyset$.*

(3) *If f is subdifferentially regular at $\bar x$, g is strictly differentiable at $\bar x$, C is convex, and K-T $(\bar x) \neq \emptyset$, then $0 \in \partial P_\alpha(\bar x)$ for all $\alpha >$ $\mathrm{dist}_0(0|K\text{-}T(\bar x))$ (or $\alpha \geq$ $\mathrm{dist}_0(0|K\text{-}T(\bar x))$ if Y^* is separable).*

Proof. (1) This follows directly from [68, Thms. 2 and 3] and inclusion (2.3).

(2) This follows from [66, Cor. 5.2.3].

(3) The proof is by [68, Thms. 2 and 3],

$$\alpha\partial P_\alpha(\bar x) = \partial f(\bar x) + \alpha g_s'(\bar x)^*\partial\ [\mathrm{dist}(\cdot|C)](g(\bar x)).$$

Let $y \in K - T(\bar x)$ be such that $\|y\|_0 < \alpha$. If Y^* is separable we can choose $\|y\|_0 = \alpha$ [21], [74]. Then $0 \in \partial P_\alpha(\bar x)$ since, by (3.5),

$$\partial\ [\mathrm{dist}(\cdot|C)](g(\bar x)) = \mathbb{B}^0 \cap N(g(\bar x)|C).$$

Remark. Proposition 2.6 extends similar results found in Garcia-Palomares [31, §4], Han and Mangasarian [33, §4], Lasserre [44], Polak, Mayne, and Wardi [59, §3], and Rosenberg [71]. All of these results apply to the finite-dimensional case with $C := \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$. They obtain results for other norms by appealing to the equivalence of norms in finite-dimensions.

It is well known that K-T $(\bar x)$ may be empty even if $\bar x$ is a local solution to $\mathcal{P}$. Nevertheless, more general multiplier rules can be established in this case. The such result is attributed to John [42]. In the next section, we generalize this result to $\mathcal{P}$.

3. A John type multiplier rule for $\mathcal{P}$. In this section we consider the problem $\mathcal{P}$ with $f := f_0 + \psi(\cdot|S)$, $f_0 : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ locally Lipschitz, $S \subset \mathbb{R}^n$ nonempty and closed, and $C \subset \mathbb{R}^m$ nonempty, closed, and convex, and derive a multiplier rule that does not depend on calmness. For this purpose let $\bar x$ be a local solution of radius ε to $\mathcal{P}$, and for each $\delta \geq 0$, consider the function $\theta_\delta : \mathbb{R} \times \mathbb{R}^n \to \overline{\mathbb{R}}$ given by

$$\theta_\delta(x) := \mathrm{dist}[(f_0(x), g(x))|C_{\bar x,\delta}] + \psi(\cdot|S \cap (\bar x + \varepsilon\mathbb{B})), \tag{3.1}$$

where

$$C_{\overline{x},\delta} := (f_0(\overline{x}) - \delta + \mathbb{R}_-) \times C \subset \mathbb{R} \times \mathbb{R}^m. \tag{3.2}$$

It is assumed that the norm chosen for $\mathbb{R} \times \mathbb{R}^m$ is such that $\|(\xi, 0)\| = |\xi|$. Observe that for each $\delta \geq 0$ we have

$$\theta_\delta(\overline{x}) \leq \delta + \inf\theta_\delta(x) \tag{3.3}$$

and if $\delta > 0$, then

$$0 < \inf\theta_\delta(x) < +\infty. \tag{3.4}$$

Thus, in particular, $\overline{x}$ is a global minimum for θ_0. The function θ_0 is a kind of exact penalty function for $\mathcal{P}$. It is similar to the Eremin–Zangwill penalty functions except that no a priori assumptions are required for $\overline{x}$ to be a global minimum for θ_0. Exact penalty functions of this type were considered by Morrison [55] in the case where $C = \{0\}$, $S = \mathbb{R}^n$, and where $\mathbb{R}^m$ is given the Euclidean norm. In this setting, Morrison showed how we can apply the methods of nonlinear least squares to solve $\mathcal{P}$. Further discussion of these penalty functions is given in Fletcher [27].

By applying the appropriate rules of the subdifferential calculus to θ_0, we can obtain a multiplier rule for $\mathcal{P}$. Unfortunately, such a direct application yields a rather uninteresting multiplier rule because of the nature of the subdifferential of the distance function $\text{dist}[\cdot|C_{\overline{x},\varepsilon}]$.

PROPOSITION 3.1. *Let Γ be a nonempty, closed, convex subset of a normed linear space X. Then* $\text{dist}(y|\Gamma)$ *is a convex function whose subdifferential is*

$$\partial \text{ dist}(y|\Gamma) := \begin{cases} \mathbb{B}^0 \cap N(y|\Gamma), & \textit{if } y \in \Gamma \\ (\text{bdry } \mathbb{B}^0) \cap N(y|\Gamma + \text{ dist}(y|\Gamma)\mathbb{B}), & \textit{otherwise.} \end{cases} \tag{3.5}$$

If Γ is not assumed to be convex, then

$$N(y|\Gamma) = \text{cl}[\cup_{\lambda \geq 0} \lambda\partial \text{ dist}(y|\Gamma)]. \tag{3.6}$$

Proof. In the convex case with $y \in \Gamma$, the formula $\partial \text{ dist}(y|\Gamma) = \mathbb{B}^0 \cap N(y|\Gamma)$ is elementary and well known. When Γ is convex and $y \notin \Gamma$, the formula is derived in Burke [9, §2]. The final formula (3.6) is due to Clarke [15, Prop. 2.4.2]. □

Thus a direct application of the chain rule [15, Thm. 2.3.10] to θ_0 would yield, according to Proposition 3.1, the trivial inclusion

$$0 \in \mathbb{R}_+ \partial f_0(\overline{x}) + \partial g(\overline{x})^* N(g(\overline{x})|C) + N(\overline{x}|S).$$

This is the reason for including the perturbation δ in definition (3.1). Due to inequality (3.3) we can apply Ekeland's variational principle [22] to obtain, for each $\delta > 0$, the existence of an $x_\delta \in (\overline{x} + \varepsilon\mathbb{B}) \cap S$ satisfying

$$\|\overline{x} - x_\delta\| \leq \sqrt{\delta},$$

and

$$\theta_\delta(x) + \sqrt{\delta}\|x - x_\delta\| > \theta_\delta(x_\delta)$$

for all $x \neq x_\delta$. Hence $\overline{x}$ is a strict global minimum of the function

$$\widehat{\theta}_\delta(x) := \theta_\delta(x) + \sqrt{\delta}\|x - x_\delta\|.$$

Now, by (3.4), we have $\theta_\delta(x_\delta) > 0$ and so $(f_0(x_\delta), g(x_\delta)) \notin C_{\overline{x},\delta}$. Thus, when we apply the appropriate rules of the subdifferential calculus (Rockafellar [66, Cor. 5.2.3]) to the inclusion $0 \in \partial\widehat{\theta}_\delta(x_\delta)$, we obtain the existence of an $x_\delta \geq 0$ and

$$y_\delta \in N(g(x_\delta)|C + \text{ dist}(g(x_\delta)|C)\mathbb{B})$$

with $\|(\lambda_\delta, y_\delta)\|_0 = 1$ such that

$$0 \in \lambda_\delta\partial f_0(x_\delta) + \partial g(x_\delta)^* y_\delta + N(x_\delta|S) + \sqrt{\delta}\mathbb{B}^0$$

for all δ with $\sqrt{\delta} < \varepsilon$. Consequently, any cluster point $(\overline{\lambda}, \overline{y})$ of $\{(\lambda_\delta, y_\delta)\}$ as $\delta \downarrow 0$ must satisfy

$$\|(\overline{\lambda}, \overline{y})\|_0 = 1, \tag{3.7}$$

$$\overline{\lambda} \geq 0,\ \overline{y} \in N(g(\overline{x})|C), \tag{3.8}$$

and

$$0 \in \overline{\lambda}\partial f_0(\overline{x}) + \partial g(\overline{x})^*\overline{y} + N(\overline{x}|S). \tag{3.9}$$

We have just proved the following theorem.

THEOREM 3.1. *Let f, g, s, C, and $\overline{x}$ be as given at the beginning of this section. Then there exist multipliers $\overline{\lambda} \geq 0$ and $\overline{y} \in N(g(\overline{x})|C)$ such that* (3.7)–(3.9) *hold.*

With a bit of work this result can be obtained from Clarke [15, Thm. 6.1.1]. Moreover, the proof that we provide has a certain similarity to Clarke's proof. We included this proof since it is simpler and more direct. Furthermore, it illustrates the intimate relationship between the multipliers and the subgradient of the distance function at $(f_0(\overline{x}), g(\overline{x}))$.

Note that if the multiplier $\overline{\lambda}$ in (3.7) is nonzero, then $\overline{\lambda}^{-1}\overline{y} \in \text{K-T}(\overline{x})$, i.e.,

$$\text{K-T}(\overline{x}) := \{\overline{\lambda}^{-1}\overline{y} : (\overline{\lambda}, \overline{y}) \text{ satisfy } (3.7)\text{–}(3.9) \text{ with } \overline{\lambda} \neq 0\}. \tag{3.10}$$

Moreover, if f and g satisfy the conditions of part (2) of Proposition 2.6, then

$$0 \in \partial P_{\overline{\lambda}^{-1}}(\overline{x}).$$

Thus the magnitude of $\overline{\lambda}$ is inversely related to the magnitude of an exact penalty parameter for $\mathcal{P}$. The multipliers $(\overline{\lambda}, \overline{y})$, for which $\overline{\lambda} = 0$, are of great significance in the analysis of $\mathcal{P}$. We call these multipliers *Fritz John multipliers* and denote them by

$$\begin{aligned} FJ(x) : &= \{\mu\overline{y} : \mu \geq 0, (0, \overline{y}) \text{ satisfies } (3.7)\text{—}(3.9)\ \} \\ &= \ker([\partial g(x)^T, I]) \cap (N(g(x)|C) \times N(x|S)), \end{aligned}$$

where

$$\ker([\partial g(x)^T, I]) := \{(y, z) \in \mathbb{R}^m \times \mathbb{R}^n : 0 \in \partial g(x)^T y + z\}.$$

Observe that $FJ(x)$ is a nonempty, closed, and convex cone for every $x \in S$ with $g(x) \in C$. Moreover, if K-T $(x) \neq \emptyset$, and g is strictly differentiable at x, then $FJ(x) = \text{rec}(\text{K-T}(x))$. Clarke [15] refers to the Kuhn–Tucker and Fritz John multipliers as the normal and abnormal multipliers, respectively.

According to Theorem 3.1, one is guaranteed of the existence of Kuhn–Tucker multipliers at a local solution $\overline{x}$ to $\mathcal{P}$ if $FJ(\overline{x}) = \{0\}$, or equivalently, if

$$\ker[\partial g(\overline{x})^T, I] \cap (N(g(x)|C) \times N(x(S)) = \{0\}.$$

This condition is precisely the constraint qualification (2.8) and (2.9) of the previous section. Thus we see that condition (2.8) is truly a fundamental property for constrained optimization. It is a natural condition under which we obtain both constraint regularity and the the existence of Kuhn–Tucker multipliers. For this reason, we will refer to (2.8) as the *basic constraint qualification* throughout the remainder of the paper.

PROPOSITION 3.2. *Let f, g, and C be as given in the beginning of this section and let $x \in S$ be such that $g(x) \in C$ and K-T $(x) \neq \emptyset$.*

(1) *If the basic constraint qualification* (2.8) *is satisfied at x, then K-T (x) is compact.*

(2) *If g is strictly differentiable at x, then* K-T (x) *is convex and* $\text{rec}(K\text{-}T(x)) = FJ(x)$, *in which case* K-T (x) *is compact if and only if the basic constraint qualification* (2.8) *is satisfied at x.*

(3) *If $\overline{x}$ is a local solution to $\mathcal{P}$ at which the basic constraint qualification* (2.8) *is satisfied, then* K-T $(\overline{x})$ *is nonempty.*

Proof. (1) If K-T (x) is not compact, then it contains an unbounded sequence $\{y_i\} \subset N(g(x)|C)$. For each $i = 1, 2, \cdots$, there exists vectors $v_i \in \partial f_0(x)$ and $z_i \in N(x|S)$ and a matrix $J_i \in \partial g(x)$ such that

$$0 = v_i + J_i^T y_i + z_i.$$

With no loss in generality, we can assume that $(y_i, z_i)(\|y_i\| + \|z_i\|)^{-1} \to (\overline{y}, \overline{z})$ and $J_i \to \overline{J}$ with $\|\overline{y}\| + \|\overline{z}\| = 1$, $\overline{y} \in N(g(x)|C), \overline{z} \in N(x|S)$, and $\overline{J} \in \partial g(x)$. But then $0 = \overline{J}^T \overline{y} + \overline{z}$ so that $FJ(x) \neq \{0\}$, a contradiction. Hence K-T (x) is compact.

(2) The convexity of K-T (x) and the equivalence $\text{rec}(\text{K-T}(x)) = FJ(x)$ follow directly from the definitions. Thus the equivalence of (2.8) with the compactness of K-T (x) follows immediately from [70, Thm. 8.4].

(3) This follows from the preceding discussion. □

Remark. Proposition 3.2 extends a well-known result of Gauvin [32]. Another generalization of Gauvin's result is obtained in Nguyen, Strodiot, and Mifflin [56], where it is assumed that $C := \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$ and that the s components of g are Lipschitz.

4. Second-order optimality conditions for $\mathcal{P}$. The second-order results of this section are based on the second-order theory for convex composite optimization developed in Burke [10] and Burke and Poliquin [11]. If $f := f_0 + \psi(\cdot|S)$ with $f_0 : \mathbb{R}^n \to \mathbb{R}$ and the sets $S \subset \mathbb{R}^n$ and $C \subset \mathbb{R}^m$ are taken to be nonempty, closed and convex, then the exact penalty functions P_α and θ_0 defined in §§2 and 3, respectively, are convex composite functions. Thus we can apply the results of [10], [11] directly to these functions. The theorems obtained in this way are very much in the spirit of those established in Levitin, Miljutin, and Osinolovski [45], Ioffe [39]–[41], Ben–Isreal, Ben-Tal, and Zlobec [4], and Rockafellar [63]–[64]. These results are distinguished by their use of the entire set of multipliers rather than a single vector of multipliers as is the case in the classical theory of second-order optimality conditions (e.g., see Hestenes [35]–[36], Pennisi [57], and Fiacco and McCormick [25]). Let us begin by reviewing the pertinent results in [66] and [11].

THEOREM 4.1 (Rockafellar [66, Cor. 5.2.3]). *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is lower semicontinuous, $g : \mathbb{R}^n \to \mathbb{R}^n$ is locally Lipschitz, and $\overline{x} \in \mathrm{dom}(q)$, where $q(x) := f(x) + h(g(x))$, is such that* (2.9) *holds. If $\overline{x}$ is a local minimum of q, then the set of multipliers*

$$(4.1)\qquad M_Q(\overline{x}) := \{y \in \partial h(\cdot)(g(\overline{x})) : 0 \in \partial f(\overline{x}) + \partial g(\overline{x})^* y\}$$

is nonempty.

THEOREM 4.2 (Burke and Poliquin [11, Thm. 4.2]). *Let $\overline{x} \in S \subset \mathbb{R}^n$ be such that $f_0 : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are twice continuously differentiable near $\overline{x}$. Moreover, let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be lower semicontinuous and convex with $g(\overline{x}) \in \mathrm{dom}(h)$, and suppose that S is closed and convex. Set $q := f_0 + \psi(\cdot|S) + h \circ g$.*

(1) *If $\overline{x}$ is a local minimum for q at which the basic constraint qualification* (2.7) *is satisfied, then $M_Q(\overline{x}) \neq \emptyset$ and*

$$(4.2)\qquad \max\{d^T(\nabla^2 f_0(\overline{x}) + \nabla^2_{xx}(\langle y, g(\overline{x})\rangle))d : y \in M_Q(\overline{x})\} \geq 0$$

for all $d \in \mathrm{cl}(K_Q(\overline{x})) \cap T(\overline{x}|S)$

$$(4.3)\qquad K_Q(x) := \{d \in \mathbb{R}^n : \exists \overline{t} > 0 \quad \textit{such that } h(g(x) + tg'(x)d) \leq h(g(x)) \forall t \in (0, \overline{t})\}.$$

(2) *If $M_Q(\overline{x}) \neq \emptyset$ and*

$$(4.4)\qquad \sup\{d^T(\nabla^2 f_0(\overline{x}) + \nabla^2_{xx}(\langle y, g(\overline{x})\rangle))d : y \in M_Q(\overline{x})\} > 0$$

for all $d \in D_Q(\overline{x})\backslash\{0\}$ where

$$(4.5)\qquad D_Q(x) := \{d \in \mathbb{R}^n : q'(x; d) \leq 0\},$$

then there is a $\gamma > 0$ such that

$$q(x) \geq q(\overline{x}) + \gamma\|x - \overline{x}\|^2$$

for all x near $\overline{x}$.

PROPOSITION 4.3. *Let $\overline{x} \in S \subset \mathbb{R}^n$ be such that $f_0 : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are continuously differentiable near $\overline{x}$. Moreover, let $h : \mathbb{R}^n \to \overline{\mathbb{R}}$ be lower semicontinuous and convex with $g(\overline{x}) \in \mathrm{dom}(h)$, suppose that S is closed, and set $q := f_0 + \varphi(\cdot|S) + h \circ g$. If*

$$(4.6)\qquad \mathrm{ran}\begin{bmatrix} g'(\overline{x}) \\ I \end{bmatrix} \bigcap \mathrm{ri}[T(g(\overline{x})|\mathrm{lev}_h(g(x))) \times T(\overline{x}|S)] \neq \emptyset$$

and

$$(4.7)\qquad \mathrm{cone}(\partial h(\cdot)(g(\overline{x}))) = N(g(\overline{x})|\mathrm{lev}_h(g(\overline{x}))),$$

then $D_Q(\overline{x}) = \overline{K_Q(\overline{x})} \cap T(\overline{x}|S)$. *Moreover, if* $0 \notin \partial h(\cdot)(g(\overline{x}))$, *then* (4.7) *is satisfied, and if the basic constraint qualification* (2.8) *holds, then* (4.6) *is satisfied.*

Proof. All but the very last statement is established in [11, Prop. 5.1]. For the last statement, we take polars in (2.7) to obtain

$$\operatorname{ran}\begin{bmatrix} g'(\overline{x}) \\ I \end{bmatrix} + (T(g(\overline{x})|\mathrm{lev}_h(g(\overline{x}))) \times T(\overline{x}|S)) = \mathbb{R}^m.$$

Now, for any subspace W and closed convex cone K the condition $W + K = \mathbb{R}^m$ implies that $W \cap \mathrm{ri}(K) \neq \emptyset$ by a simple separation argument. This establishes the result. □

We now apply these results to $\mathcal{P}$. The result is a sufficiency theorem which does not require a constraint qualification. The result is obtained by applying Theorem 4.2 to the function θ_0.

THEOREM 4.4. *Let* S *and* C *be nonempty closed convex subsets of* $\mathbb{R}^n$ *and* $\mathbb{R}^m$, *respectively, and let* $\overline{x} \in S$ *be such that* $f_0 : \mathbb{R}^n \to \mathbb{R}$ *and* $g : \mathbb{R}^n \to \mathbb{R}^m$ *are twice continuously differentiable near* $\overline{x}$ *and* $g(\overline{x}) \in C$. *Set* $f := f_0 + \psi(\cdot|S)$ *and consider the problem* $\mathcal{P}$. *If the set of multipliers*

$$M_{\mathcal{P}}(\overline{x}) := \{(\lambda, y) \in \mathbb{R} \times \mathbb{R}^m : (3.7)\text{–}(3.9) \textit{ are satisfied}\} \tag{4.8}$$

is nonempty and

$$\max\{d^T(\lambda\nabla^2 f_0(\overline{x}) + \nabla^2_{xx}(\langle y, g(\overline{x})\rangle))d : (\lambda, y) \in M_{\mathcal{P}}(\overline{x})\} > 0 \tag{4.9}$$

for every $d \in D_{\mathcal{P}}(\overline{x})$ *where*

$$D_{\mathcal{P}}(\overline{x}) := \{d \in T(\overline{x}|S) : \nabla f_0(\overline{x})^T d \leq 0, g'(\overline{x})d \in T(g(\overline{x})|C)\}, \tag{4.10}$$

then there is a $\gamma > 0$ *such that*

$$f_0(x) \geq f_0(\overline{x}) + \gamma\|x - \overline{x}\|^2 \tag{4.11}$$

for every $x \in \Omega := \{x \in S : g(x) \in C\}$ *near* $\overline{x}$.

Proof. Consider part (2) of Theorem 4.2 as it applies to the function θ_0 defined in (3.1). We begin by defining the functions f_0, g, and h and the set S that appear in Theorem 4.2. For the sake of clarity, we denote these functions and set as $f_{0_{(4.2)}}$, $g_{(4.2)}$, $h_{(4.2)}$, and $S_{(4.2)}$, respectively. For the remainder of the proof the functions f_0 and g, and the set S, will refer to those that are given in the statement of Theorem 4.4. With this notation we define $f_{0_{(4.2)}} \equiv 0$, $g_{(4.2)} := (f_0, g)$, $h_{(4.2)} :=$ $\mathrm{dist}[\cdot|C_{\overline{x},0}]$, and $S_{(4.2)} := S \cap (\overline{x} + \varepsilon\mathbb{B})$. The set $M_Q(\overline{x})$ is given by

$$\{(\lambda, y) : \|(\lambda, y)\|_0 \leq 1 \text{ and } (3.8)\text{–}(3.9) \text{ hold for } f_{0_{(4.2)}} \text{ and } g_{(4.2)}\} \supset M_{\mathcal{P}}(\overline{x}),$$

and the set $D_Q(\overline{x})$ is given by

$$\begin{aligned}
&\{d : \psi^*(d|N(\overline{x}|S)) + \psi^*(d|[0,1]\nabla f_0(x) + g'(x)^T(\mathbb{B}^0 \cap N(g(x)|C))) \leq 0\} \\
=&\{d \in T(\overline{x}|S) : \lambda\nabla f_0(\overline{x})^T d + y^T g'(\overline{x})d \leq 0 \ \forall\lambda \in [0,1], y \in \mathbb{B}^0 \cap N(g(\overline{x})|C)\} \\
=&\{d \in T(\overline{x}|S) : \nabla f_0(\overline{x})^T d \leq 0, g'(\overline{x})d \in T(g(\overline{x})|C)\} \\
=&D_{\mathcal{P}}(\overline{x}),
\end{aligned}$$

where the line follows by choosing the norm on $\mathbb{R} \times \mathbb{R}^m$ to be $|\xi| + \|y\|_0$ for every $(\xi, y) \in \mathbb{R} \times \mathbb{R}^m$. Since inequality (4.9) implies inequality (4.4), we have the existence of $\gamma > 0$ such that

$$\theta_0(x) \geq \theta_0(\overline{x}) + \gamma\|x - \overline{x}\|^2 \tag{4.12}$$

for all x near $\overline{x}$, where by the theorem is proved. □

Remarks. (1) The theorem actually establishes inequality (4.12), which is stronger than inequality (4.11).

(2) We could just as well have used the multiplier set $M_{\mathcal{Q}}(\overline{x})$ in (4.9), but, since the maximum is positive, both of these multiplier sets yield the same value in (4.9).

Unfortunately, without a constraint qualification, the same trick cannot be applied to obtain a second-order necessary condition for $\mathcal{P}$. The problem is that $M_{\mathcal{P}}(\overline{x}) \subsetneq M_{\mathcal{Q}}(\overline{x})$ with $(0,0) \in M_{\mathcal{Q}}(\overline{x})$. Consequently (4.2) is valid for all $d \in \mathbb{R}^n$ and it does not imply (4.9) with the weak inequality. On the other hand, if the sets C and S are polyhedral convex, then such a result can be established (e.g., see [35] and [36] or [25]).

If one is willing to assume the basic constraint qualification (2.7), then, by applying Theorem 4.2 to P_α both second-order necessary and sufficient conditions for $\mathcal{P}$ can be obtained. To establish this result, we require the following lemma.

LEMMA 4.5. *Let X and Y be normed linear spaces and let C be a nonempty closed convex subset of Y. Moreover, let $\overline{x} \in X$, $f : X \to \overline{\mathbb{R}}$, and $g : X \to Y$ be such that $\partial f(\overline{x}) \neq \emptyset$, f is subdifferentially regular at $\overline{x}$, $g(\overline{x}) \in C$, and g is strictly differentiable at $\overline{x}$. If the set* K-T $(\overline{x})$ *is nonempty, then*

$$\begin{aligned}\{d \in X : P_\alpha^0(\overline{x}; d) \leq 0\} &= \{d \in X : f^0(\overline{x}; d) = 0, g_s'(\overline{x})d \in T(g(\overline{x})|C)\} \\ &=: D_{\mathcal{P}}(\overline{x})\end{aligned}$$

for all $\alpha >$ $\operatorname{dist}_0(0|K\text{-}T(\overline{x}))$.

Proof. The hypotheses and Rockafellar [68, Thms. 2 and 3] imply that

$$\partial P_\alpha(\overline{x}) = \partial f(\overline{x}) + \alpha g_s'(\overline{x})^*(\mathbb{B}^0 \cap N(g(\overline{x})|C)).$$

Thus, if $\alpha >$ $\operatorname{dist}_0(0|\text{K-T}(\overline{x}))$, then clearly

$$D_{\mathcal{P}}(\overline{x}) \subset \{d \in X : P_\alpha^0(\overline{x}; d) \leq 0\}.$$

On the other hand, let $d \in \{d : P_\alpha^0(\overline{x}; d) \leq 0\}$. Then for each $y \in$ K-T $(\overline{x})$ with $\|y\|_0 < \alpha$ there is a $z \in \partial f(\overline{x})$ such that $0 = z + g_s'(\overline{x})^* y$. Hence,

$$\begin{aligned}0 &\geq P_\alpha^0(\overline{x}; d) \\ &\geq \langle z + \alpha g_s'(\overline{x})^* \frac{y}{\|y\|_0}, d\rangle \\ &= (1 - \frac{\alpha}{\|y\|_0})\langle z, d\rangle,\end{aligned}$$

and so $f^0(\overline{x}; d) \geq \langle z, d\rangle \geq 0$. But $0 \in N(g(\overline{x})|C) \cap \mathbb{B}^0$ so that $f^0(\overline{x}; d) \leq 0$. Consequently, $f^0(\overline{x}; d) = 0$ and $g_s'(\overline{x})d \in T(g(\overline{x})|\mathbb{B})$. □

Remark. The set $D_{\mathcal{P}}(\overline{x})$ given above is the obvious generalization of the set defined in (4.10) to which it reduces under the hypotheses of Theorem 4.4.

THEOREM 4.6. *Let S and C be nonempty closed convex subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively, and let $\overline{x} \in S$ be such that $f_0 : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are twice continuously differentiable near $\overline{x}$, $g(\overline{x}) \in C$, and the basic constraint qualification (2.7) holds. Set $f := f_0 + \psi(\cdot|S)$ and consider the problem $\mathcal{P}$.*

(1) *If $\overline{x}$ is a local solution to $\mathcal{P}$, then*

$$\max\{d^T(\nabla^2 f_0(\overline{x}) + \nabla^2_{xx}(\langle y, g(\overline{x})\rangle))d : y \in K\text{-}T(\overline{x})\} \geq 0 \tag{4.13}$$

for all $d \in D_{\mathcal{P}}(\overline{x})$.

(2) *If K-T $(\overline{x}) \neq \emptyset$ and*

$$\max\{d^T(\nabla^2 f_0(x) + \nabla^2_{xx}(\langle y, g(\overline{x})\rangle)))d : y \in K\text{-}T(\overline{x})\} > 0 \tag{4.14}$$

for all $d \in D_{\mathcal{P}}(\overline{x})\backslash\{0\}$, then for each

$$\alpha > \overline{\alpha} := \max\{\|y\|_0 : y \in K\text{-}T(\overline{x})\} \tag{4.15}$$

there are scalars $\varepsilon > 0$ and $\gamma > 0$ such that

$$P_\alpha(x) \geq P_\alpha(\overline{x}) + \gamma\|x - \overline{x}\|^2$$

for all $x \in \overline{x} + \varepsilon\mathbb{B}$, and

$$f_0(x) \geq f_0(\overline{x}) + \gamma\|x - \overline{x}\|^2$$

for all $x \in \Omega \cap (\overline{x} + \varepsilon\mathbb{R})$ where

$$\Omega := \{x \in S : g(x) \in C\}.$$

Proof. In $\mathcal{Q}$ take $h := \alpha \operatorname{dist}(\cdot|C)$. Then, by Proposition 4.3 and Lemma 4.5,

$$D_{\mathcal{Q}}(\overline{x}) = \operatorname{cl}(K_{\mathcal{Q}}(\overline{x})) \cap T(\overline{x}|S) = D_{\mathcal{P}}(\overline{x}) \tag{4.16}$$

as long as $\alpha > \|y\|_0$ for some $y \in$ K-T$(\overline{x})$. Moreover, by part (1) of Proposition 3.2, the set K-T $(\overline{x})$ is compact. Hence if K-T $(\overline{x}) \neq \emptyset$, then $\overline{\alpha}$ is finite and for any $\alpha \geq \overline{\alpha}$ one has

$$M_{\mathcal{Q}}(\overline{x}) = \text{K-T}(\overline{x}). \tag{4.17}$$

(1) By Corollary 2.5.1, there is an $\alpha > 0$ such that $\overline{x}$ is a local minimum for P_α. Taking $h := \alpha \operatorname{dist}(\cdot|C)$ in $\mathcal{Q}$, we get $M_{\mathcal{Q}}(\overline{x}) \subset$ K-T$(\overline{x})$, where α is chosen so that $\alpha >$ $\operatorname{dist}_0(0|\text{K-T}(\overline{x}))$. The result then follows from (4.16) and part (1) of Theorem 4.2.

(2) By taking $\alpha > 0$ to satisfy (4.15) and by observing (4.16) and (4.17), the result is an immediate consequence of part (2) of Theorem 4.2 with $h := \alpha \operatorname{dist}(\cdot|C)$. □

Remark. For the case in which C and S are polyhedral convex, Theorem 4.6 is also obtained by Ioffe [39]–[41], Ben–Israel, Ben–Tal, and Zlobec [4], and Rockafellar [63]–[64].

In Theorem 4.6 we obtain second-order necessary and sufficient conditions for $\mathcal{P}$ from the corresponding second-order conditions for P_α. This approach is the reverse of that which is usually taken in the literature. In particular, Charalambous [13, Thm. 2], Han and Mangasarian [33, Thm. 4.6], and Lasserre [44, Thm. 2] essentially show that if K-T $(\bar{x}) \neq \emptyset$ and the second-order sufficiency condition of Pennisi [57, Thm. 3.3] holds for some $y \in$ K-T$(\bar{x})$, then $\bar{x}$ is a strict local minimum for P_α for all $\alpha > \|y\|_0$. These results do not require the imposition of the basic constraint qualification (2.7). On the other hand, they do require the application of a stronger second-order sufficiency condition. In the next result, we obtain a result, paralleling those of Charalambous, Han and Mangasarian, and Lasserre.

THEOREM 4.7. *Let S, C, $\bar{x}$, f_0, and g be as in the statement of Theorem* 4.6, *except that the basic constraint qualification may fail to hold at $\bar{x}$. If there exists $\bar{y} \in K\text{-}T(\bar{x})$ such that*

$$d^T(\nabla^2 f_0(\bar{x}) + \nabla^2_{xx}(\langle \bar{y}, g(\bar{x})\rangle))d > 0$$

for every $d \in D_{\mathcal{P}}(\bar{x})\backslash\{0\}$, then for each $\alpha > \|y\|_0$ there are scalars $\varepsilon > 0$ and $\gamma > 0$ such that

$$P_\alpha(x) \geq P_\alpha(\bar{x}) + \gamma\|x - \bar{x}\|^2$$

for all $x \in \bar{x} + \varepsilon\mathbb{B}$ and

$$f_0(x) \geq f_0(\bar{x}) + \gamma\|x - \bar{x}\|^2$$

for all $x \in \Omega \cap (\bar{x} + \varepsilon\mathbb{B})$.

Proof. For this choice of α (4.16) holds and by part (3) of Proposition 2.6, $0 \in \partial P_\alpha(\bar{x})$ with $y \in M_{\mathcal{Q}}(\bar{x})$ where $h := \alpha \operatorname{dist}(\cdot|C)$. Hence, the result again follows directly from part (2) of Theorem 4.2. □

Before leaving this section we obtain yet another sufficiency result for $\mathcal{P}$. It is a first-order sufficiency result and is a direct consequence of Lemma 4.5. The result is similar to results by Howe [37], Rosenberg [71, Thm. 3], and Bazaraa and Goode [3, Thms. 2.1, 2.2, 3.1, and 4.1].

THEOREM 4.8. *Let X, Y, $\bar{x}$, f, and g be as in the statement of Lemma* 4.5 *where it is further assumed that X is finite-dimensional. If the set $K\text{-}T$ $(\bar{x})$ is nonempty and $D_{\mathcal{P}}(\bar{x}) = \{0\}$, then there are scalars $\varepsilon > 0$ and $\gamma > 0$ such that*

$$P_\alpha(x) \geq P_\alpha(\bar{x}) + \gamma\|x - \bar{x}\|$$

for all $x \in (\bar{x} + \varepsilon\mathbb{B})$ and $\alpha >$ $\operatorname{dist}_0(0|K\text{-}T(\bar{x}))$, and

$$f(x) \geq f(\bar{x}) + \gamma\|x - \bar{x}\|$$

for all $x \in \{x : g(x) \in C\} \cap (\bar{x} + \varepsilon\mathbb{B})$.

Proof. From Lemma 4.5, $P^0_\alpha(\bar{x}; d) > 0$ for all $d \neq 0$. By Rockafellar [68, Thms. 2 and 3], $P^0_\alpha(\bar{x}; d) = P'_\alpha(\bar{x}; d)$. The result now easily follows with $\gamma = \inf\{P'_\alpha(\bar{x}; d) : \|d\| = 1\} > 0$. □

5. Convex programming. Eremin and Zangwill originated the study of exact penalization in the context of convex programming. In this section, we extend this theory to the problem $\mathcal{P}$. The step in this process is to establish an equivalence between the problem $\mathcal{P}$ and a problem $\widetilde{\mathcal{P}}$ to which the classical theory of convex programming applies [35], [36], [38], [43], [46], [48], [51], [70], [72]. To this end, let X be a real normed linear space, Y a real reflexive Banach space, $C \subset X$ and $S \subset X$ be nonempty, closed, and convex, and set

$$\widetilde{C} := \operatorname{cl}\{(\lambda, \lambda y) : \lambda \geq 0, y \in C\},$$

where the closure is taken with respect to the product topology on $\mathbb{R} \times Y$.

Given $g : X \to Y$ we define $G : X \to \mathbb{R} \times Y$ by

$$G(x) := (-1, -g(x))$$

for all $x \in X$. Consider the constrained optimization problem

$$(\widetilde{\mathcal{P}}) \qquad \begin{array}{l} \text{minimize } f(x) \\ \text{subject to } G(x) \le 0, \end{array}$$

where $f := f_0 + \psi(\cdot|S)$ with $f_0 : X \to \mathbb{R}$ a convex function and where "$\le$" denotes the partial order induced on $\mathbb{R} \times Y$ by $\widetilde{C}$, i.e., $\underset{\sim}{y_1} \le y_2$ if and only if $y_2 - y_1 \in \widetilde{C}$. Observe that $x \in X$ solves $\mathcal{P}$ if and only if x solves $\widetilde{\mathcal{P}}$. We now develop a purely convex theory for $\mathcal{P}$ based upon that which already exists for $\widetilde{\mathcal{P}}$.

LEMMA 5.1. *Let $G : X \to \mathbb{R} \times Y$ and $\widetilde{C} \subset \mathbb{R} \times Y$ be as given above. Then the following conditions are equivalent.*

(1) *G is convex with respect to $\widetilde{C}$; i.e., $G(\lambda x + (1 - \lambda)y) \le \lambda G(x) + (1 - \lambda)G(y)$ for every $x, y \in X$ and $\lambda \in [0, 1]$.*

(2) *g is concave with respect to $\operatorname{rec}(C)$; i.e., $g(\lambda x + (1-\lambda)y) - [\lambda g(x) + (1-\lambda)g(y)] \in \operatorname{rec}(C)$ for every $x, y \in X$ and $\lambda \in [0, 1]$.*

(3) *For each $y \in \operatorname{bar}(C)$ the mapping $g_y : X \to \mathbb{R}$, given by $g_y(\cdot) := \langle y, g(\cdot)\rangle$, is convex.*

Moreover, each of the above conditions imply that the distance function $\operatorname{dist}(g(\cdot)|C)$ is convex.

Proof. (1) $\iff$ (2): Let $x_1, x_2 \in X$ and choose $\lambda \in [0, 1]$. Then G is convex with respect to $\widetilde{C}$ if and only if

$$\lambda(-1, -g(x_1)) + (1 - \lambda)(-1, -g(x_2)) - (-1, -g(\lambda x_1 + (1 - \lambda)x_2)) \in \widetilde{C},$$

or equivalently

$$g(\lambda x_1 + (1 - \lambda)x_2) - [\lambda g(x_1) + (1 - \lambda)g(x_2)] \in \operatorname{rec}(C)$$

since $\operatorname{rec}(C) = \{y : (0, y) \in \widetilde{C}\}$. This is equivalent to saying that g is concave with respect to $\operatorname{rec}(C)$.

(2) $\iff$ (3): The mapping g is concave with respect to $\operatorname{rec}(C)$ if and only if for every $x_1, x_2 \in X$ and $\lambda \in [0, 1]$

$$g(\lambda x_1 + (1 - \lambda)x_2) - [\lambda g(x_1) + (1 - \lambda)g(x_2)] \in (\operatorname{bar}(C))^0$$

since $[\operatorname{bar}(C)]^0 = \operatorname{rec}(C)$. This is equivalent to saying that

$$\langle y, g(\lambda x_1 + (1 - \lambda)x_2)\rangle \le \langle y, \lambda g(x_1) + (1 - \lambda)g(x_2)\rangle$$

for every $x_1, x_2 \in X$, $\lambda \in [0, 1]$, and $y \in \operatorname{bar}(C)$; i.e., $\langle y, g(\cdot)\rangle$ is convex for every $y \in \operatorname{bar}(C)$.

Finally, if any one of (1)–(3) hold, then clearly (3) is valid. Hence, for every $y \in \operatorname{bar}(C) = \operatorname{dom}(\psi^*(\cdot|C))$ the function

$$\langle y, g(\cdot)\rangle - \psi^*(y|C)$$

is convex. Therefore,

$$\operatorname{dist}(g(\cdot)|C) := \sup\{\langle y, g(\cdot)\rangle - \psi^*(y|C) : y \in \mathbb{B}^0\}$$

is convex since it is the supremum of a collection of convex functions. □

Remark. If C is bounded, then g is concave with respect to the $\operatorname{rec}(C)$ if and only if g is affine, and if $C := \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$, then g is concave with respect to $\operatorname{rec}(C)$ if and only if g_i is convex for $i = 1, \cdots, s$ and g_i is affine for $i = s+1, \cdots, m$.

LEMMA 5.2. *Let $\widetilde{L} : S \times \widetilde{C}^* \to \mathbb{R}$ be the standard Lagrangian for $\widetilde{\mathcal{P}}$ where $\widetilde{C}^* := -\widetilde{C}^0$, i.e.,*

$$\widetilde{L}(x, z) := f(x) + \langle z, G(x)\rangle,$$

and define $L : S \times Y^* \to \mathbb{R}$ by

$$L(x, y) := f(x) + \langle y, g(x)\rangle - \psi^*(y|C).$$

Suppose that $f := f_0 + \psi(\cdot|S)$ with $f_0 : X \to \mathbb{R}$ convex and $g : X \to Y$ is concave with respect to $\operatorname{rec}(C)$ so that both $\widetilde{L}$ and L are convex–concave saddle functions by the previous lemma. Then $(x_0, (\xi_0, -y_0)) \in S \times \widetilde{C}^*$ is a saddle point for $\widetilde{L}$ if and only if (x_0, y_0) is a saddle point for L in which case $\xi_0 = \psi^*(y_0|C)$, $y_0 \in N(g(x_0)|C)$, and $g(x_0) \in C$.

Proof. By direct computation we verify that

$$\widetilde{C}^* := \{(\xi, -y) | (\xi, y) \in \operatorname{epi}(\psi^*(\cdot|C))\}.$$

If $(x_0, (\xi_0, -y_0))$ is a saddle point for $\widetilde{L}$, then, in particular, $x_0 \in S$ and

$$\langle(\xi, -y), (-1, -g(x_0))\rangle \le \langle(\xi_0, -y_0), (-1, -g(x_0))\rangle$$

for every $(\xi, y) \in \operatorname{epi}(\psi^*(\cdot|C))$, or equivalently,

$$\langle y, g(x_0)\rangle - \xi \le \langle y_0, g(x_0)\rangle - \xi_0 \tag{5.1}$$

for every $(\xi, y) \in \operatorname{epi}(\psi^*(\cdot|C))$. But this can occur if and only if

$$\xi_0 = \psi^*(y_0|C), g(x_0) \in C, \langle y_0, g(x_0)\rangle = \psi^*(y_0|C),$$

and

$$L(x_0, y) \le L(x_0, y_0)$$

for every $y \in \operatorname{bar}(C)$. To see this, set $y = y_0$ in (5.1) to get $\xi_0 = \psi^*(y_0|C)$, next set $y = y_0 + z$ in (5.1) to get $\langle z, g(x_0)\rangle \le \psi^*(z|C)$ for all $z \in \operatorname{bar}(C)$, and so $g(x_0) \in C$. Finally, having $g(x_0) \in C$ we obtain from (5.1) that

$$0 \le \langle y_0, g(x_0)\rangle - \psi^*(y_0|C) \le 0.$$

The reverse implication is obvious.

By employing the fact that $\xi_0 = \psi^*(y_0|C)$, $g(x_0) \in C$, and $\psi^*(y_0|C) = \langle y_0, g(x_0)\rangle$, we obtain from the other half of the saddle point inequalities for $\widetilde{L}$ that

$$L(x_0, y_0) = \widetilde{L}(x_0, (\xi_0, -y_0)) \le f(x) + \langle y_0, g(x)\rangle - \psi^*(y_0|C)$$

for every $x \in S$, or equivalently,

$$L(x_0, y_0) \le L(x, y_0)$$

for all $x \in S$ whereby the lemma is established. □

Having obtained the equivalence of the saddle point conditions for $\widetilde{L}$ and L, we can now simply translate the saddle point results for $\widetilde{\mathcal{P}}$ into similar results for $\mathcal{P}$. In this way, we obtain the following two results from [46, Cor. 1, p. 219] and [46, Thm. 2, p. 221], respectively.

THEOREM 5.3. *Let X be a real normed linear space and Y a real reflexive Banach space, let $S \subset X$ and $C \subset Y$ be nonempty, closed, and convex, and suppose that $f := f_0 + \psi(\cdot|S)$ with $f_0 : X \to \mathbb{R}$ convex, and $g : X \to Y$ is concave with respect to* $\mathrm{rec}(C)$, *and there is an $x \in S$ such that $g(x) \in \mathrm{int}(C)$. If $\overline{x}$ solves $\mathcal{P}$, then there is a $\overline{y} \in N(g(\overline{x})|C)$ such that $(\overline{x}, \overline{y})$ is a saddle point for $L(x, y)$.*

Remark. If X and Y are finite-dimensional, then we need only assume that there is an $x \in S$ such that $g(x) \in \mathrm{ri}(C)$.

THEOREM 5.4. *Let X, Y, S, C, g, and f be as in the statement of Theorem* 5.3. *If there exists an $\overline{x} \in S$ and $\overline{y} \in \mathrm{bar}(C)$ such that $(\overline{x}, \overline{y})$ is a saddle point for $L(x, y)$, then $\overline{x}$ solves $\mathcal{P}$.*

Further results of this type can also be obtained. Theorems 5.3 and 5.4 are presented only to give the flavor of what can be said in the convex case. In this setting, the most natural notion of a Kuhn–Tucker multiplier is derived from that of a saddle point of L. Thus, for the convex case, we extend the definition of K-T (x) as follows;

$$\text{K-T}(x) := \{y \in \mathrm{bar}(C) : (x, y) \text{ is a saddle point for } L\}.$$

Our primary result on exact penalization in the convex case now follows.

THEOREM 5.5. *Let X, Y, S, C, f, and g be as in the statement of Theorem* 5.3, *let $\overline{x} \in S$, and consider the following two conditions:*

(A) *f is continuous near $\overline{x}$ and g is strictly differentiable at $\overline{x}$.*

(B) *X and Y are finite-dimensional and g is Lipschitz near $\overline{x}$.*

The following statements are equivalent:

(1) *$\mathcal{P}$ is calm at $\overline{x}$.*

(2) *$\overline{x}$ is a global minimum of P_α for all α sufficiently large.*

Moreover, if either (A) *or* (B) *holds, then* (1) *and* (2) *are equivalent to*

(3) *K-T $(\overline{x}) \neq \emptyset$.*

Furthermore, given $\overline{y} \in$ K-T$(\overline{x})$, then $\overline{x}$ is a global minimum for P_α for all $\alpha \ge \|\overline{y}\|_0$ and if $\alpha >$ $\mathrm{dist}_0(0|\text{K-T}(\overline{x}))$, then

$$\arg\min\{P_\alpha(x) : x \in X\} \subset \arg\min\{f(x) : g(x) \in C\}.$$

Proof. By Lemma 5.1, P_α is a convex function for all $\alpha \ge 0$; consequently, any local minimum of P_α is a global minimum of P_α. Therefore, the equivalence of (1) and (2) is a consequence of Theorem 2.1.

The proof that (3) is equivalent to (1) and (2) is essentially identical under the two hypotheses (A) and (B), except that we use [66, Cor. 5.2.3] in the finite-dimensional case and [68, Thms. 2 and 3] in the infinite-dimensional case. Hence, we provide the proof only when (A) is assumed.

We begin by assuming (2) and showing that (3) holds. From (2) there is a $\overline{y} \in \partial\ \mathrm{dist}(\cdot|C)(g(\overline{x}))$ with $0 \in \partial f(\overline{x}) + \alpha g_s'(\overline{x})^*\overline{y}$, or equivalently, there is a $\overline{y} \in N(g(\overline{x})|C)$ such that

$$0 \in \partial_x L(\overline{x}, \overline{y}),$$

since $\partial\ \mathrm{dist}(\cdot|C)(y(\overline{x})) = \mathbb{B}^0 \cap N(g(\overline{x})|C)$ by Proposition 3.1. Consequently,

$$L(\overline{x}, \overline{y}) \leq L(x, \overline{y})$$

for all $x \in X$, since $L(x, \overline{y})$ is convex in x by Lemma 5.1. Finally, since

$$\begin{aligned}\psi(g(\overline{x})|C) &= \langle \overline{y}, g(\overline{x})\rangle - \psi^*(\overline{y}|C)\\ &= \sup\{\langle y, g(\overline{x})\rangle - \psi^*(y|C) : y \in Y\},\end{aligned}$$

we have that

$$L(\overline{x}, y) \leq L(\overline{x}, \overline{y})$$

for all $y \in Y$.

Next we assume that (3) holds and establish (2). Since $L(\overline{x}, y) \leq L(\overline{x}, \overline{y})$ for all $y \in Y$ we know that

$$0 = \psi(g(\overline{x})|C) = \langle \overline{y}, g(\overline{x})\rangle - \psi^*(\overline{y}|C).$$

Next, let $x \in S$ and choose $\alpha \geq \|\overline{y}\|_0$, then

$$\begin{aligned}P_\alpha(\overline{x}) &= L(\overline{x}, \overline{y})\\ &\leq L(x, \overline{y})\\ &\leq \sup\{L(x, y) : y \in \alpha\mathbb{B}\}\\ &= f(x) + \alpha\sup\{\langle y, g(x)\rangle - \psi^*(y|C) | y \in \mathbb{B}^0\}\\ &= P_\alpha(x).\end{aligned}$$

Hence $\overline{x}$ is a global minimum for $P_\alpha(x)$ for all $\alpha \geq \|\overline{y}\|_0$.

To prove the last statement of the theorem choose $\overline{y} \in$ K-T$(\overline{x})$ such that $\alpha > \|\overline{y}\|_0$. Setting $\overline{\alpha} = \|\overline{y}\|_0$, we know that $0 \in \partial P_{\overline{\alpha}}(\overline{x})$ and $0 \in \partial P_\alpha(\overline{x})$ so that $\overline{x}$ is a global minimum for both $P_{\overline{\alpha}}$ and P_α. Thus, in particular, arg $\min\{P_\alpha(x) : x \in X\} \neq \emptyset$. Let $\widetilde{x} \in$ arg $\min\{P_\alpha(x) : x \in X\}$, we need to show that $\widetilde{x} \in$ arg $\min\{f(x) : g(x) \in C\}$. For this, it is sufficient to show that $f(\widetilde{x}) \leq f(\overline{x})$ and $g(\widetilde{x}) \in C$. Due to the nature of $\overline{x}$ and $\widetilde{x}$ we have

$$f(\widetilde{x}) + \alpha\ \mathrm{dist}(g(\widetilde{x})|C) \leq f(\overline{x}) + \alpha\ \mathrm{dist}(g(\overline{x})|C)$$

and

$$f(\overline{x}) + \overline{\alpha}\ \mathrm{dist}(g(\overline{x})|C) \leq f(\widetilde{x}) + \overline{\alpha}\ \mathrm{dist}(g(\widetilde{x})|C).$$

By adding these inequalities we find that

$$(\alpha - \overline{\alpha})\ \mathrm{dist}(g(\widetilde{x})|C) \leq (\alpha - \overline{\alpha})\ \mathrm{dist}(g(\overline{x})|C).$$

Hence $g(\widetilde{x}) \in C$ and $f(\widetilde{x}) \leq f(\overline{x})$. □

Remark. The form of Theorem 5.5 is based on Rosenberg [71, Thm. 2]. This result extends similar results appearing in Eremin [23], Zangwill [75], Pietrzykowski

[58], Luenberger [47], Charalambous [13], Han and Mangasarian [33], Lasserre [44], Garcia-Palomares [31], Rosenberg [71], and Bertsekas [6].

6. Historical review.

In this section we attempt to provide a chronology of those results that establish the existence of an exact penalty parameter. We apologize for any omission or oversight.

It seems apparent that the big-M method for linear programming is the precursor of exact penalization techniques for nonlinear programming, especially since the initial results were obtained for the convex programming case. However, we are uncertain that this was indeed the motivation. Our earliest reference for the big-M method is Charnes, Cooper, and Anderson [14, §4]. The precise origins of the method are unknown to us. Our earliest reference for exact penalization in nonlinear programming is Eremin [23]. In this paper, Eremin considers the case of convex programming with $C = \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$ and $S = \mathbb{R}^n$. In [23, Thm. 2], he shows that if $\bar{y} \in \text{K-T}(\bar{x})$, then $\bar{x}$ is a global minimum for P_α whenever $\alpha > \|y\|_0$ when $\mathbb{R}^m$ is endowed with the ℓ_1 norm. At essentially the same time, Zangwill [75] published his well-known paper. Zangwill considered the case of convex programming with $C = \mathbb{R}^m_-$ and $S = \mathbb{R}^n$ and showed that if $\bar{x}$ solved $\mathcal{P}$ and $g(x_0) \in \text{int}(C)$, then $\bar{x}$ minimized P_α for all $\alpha > (f(x_0) - f(\bar{x}) + 1)(\max[g_i(x_0) : i = 1, \cdots, m])^{-1}$. This result can be used to show that K-T $(\bar{x}) \neq \emptyset$, and so is somewhat deeper than Eremin's result.

Pietrzykowski [58] provides the result for the nonconvex case. He considers the instance of $\mathcal{P}$ where $C = \mathbb{R}^s \times \{0\}_{\mathbb{R}^{m-s}}$ and $S = \mathbb{R}^n$. The analysis that Pietrzykowski gives a reminiscent of Zangwill's. He shows that if $\bar{x}$ is a strict local minimum for $\mathcal{P}$, near which f and g are differentiable and at which $g'(\bar{x})$ is surjective, then $\bar{x}$ is a strict local minimum for P_α for all α sufficiently large. Pietrzykowski's result can be used to show that K-T $(\bar{x}) \neq \emptyset$ under these hypotheses.

Luenberger [47] considers exact penalization in the setting of optimal control. We interpret his result as it applies to $\mathcal{P}$. In this context, Luenberger has $C = \mathbb{R}^m_-$ and $S = \mathbb{R}^n$ and assumes that $\bar{x}$ is a local minimum for $\mathcal{P}$ at which there exists a $\bar{y} \in N(g(\bar{x})|C)$ such that $\bar{x}$ is a local minimum for $L(x, \bar{y})$. Under these circumstances Luenberger shows that $\bar{x}$ is a local minimum for P_α for all $\alpha \geq \|\bar{y}\|_0$ where $\mathbb{R}^m$ is endowed with the ℓ_1 norm. Luenberger's proof is the same as that provided by Eremin. Clearly, Luenberger's result applies in the convex case subject to the appropriate constraint qualification, but it can also be applied to cases in which the second-order sufficiency condition of Pennisi [57] holds. Luenberger himself only states that this result applies "under standard regularity conditions."

Evans, Gould, and Tolle [24] consider the case where $C = \mathbb{R}^m_-$, $S = \mathbb{R}^n$, and f and g are continuously differentiable. In this context, their nondifferentiable exact penalty functions are quite different from the Eremin–Zangwill exact penalty functions. For these new functions they provide some exactness results that are similar in spirit to those of Eremin, Zangwill, and Pietrzykowski.

Howe [37] considers the case in which $C = \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$, $S = \mathbb{R}^n$, and f and g are continuously differentiable. His result is the appearance of the type of sufficiency result given in Theorem 4.8. He shows that if $D_{\mathcal{P}}(\bar{x}) = \{0\}$, then $\bar{x}$ is a local minimum for P_α for all α sufficiently large.

Bandler and Charalambous [2] consider the same case as Evans, Gould, and Tolle [24] and derive yet another type of nondifferentiable exact penalty function. For this exact penalty function they provide an exactness result that is similar in spirit to those of Eremin, Zangwill, and Pietrzykowski.

Bertsekas [6] investigates the case of convex programming with $C = \mathbb{R}^m_-$ and

$X = \mathbb{R}^n$, and establishes necessary and sufficient conditions for a function of the form

$$\pi(x) = f(x) + \sum_{i=1}^{m} p_i(g_i(x))$$

to be exact for $\mathcal{P}$. If $(\overline{x}, \overline{y})$ is a saddle point for $L(x, y)$ he shows that

$$\lim_{t \to 0^+} \frac{p_i(t)}{t} \geq \overline{y}^{(i)} \quad i = 1, \cdots, m,$$

with

$$\arg\min\{\pi(x)\} = \arg\min\{f(x) : g(x) \in C\},$$

if

$$\lim_{t \to 0^+} \frac{p_i(t)}{t} > \overline{y}^{(i)}.$$

We obtain Eremin's result as a special case. Bertsekas also applies his result to the exact penalty functions of Evans, Gould, and Tolle.

Charalambous [13] is the first to consider more general norms in the construction of P_α. Specifically, Charalambous considers the case $C = \mathbb{R}^m_-, S = \mathbb{R}^n$ where f and g are continuously differentiable. He then utilizes the ℓ_p-norms to form P_α. Charalambous establishes two key results. In the result, he considers the convex programming case and shows that if $(\overline{x}, \overline{y})$ is a saddle point for $L(x, y)$, then $\overline{x}$ is a global minimum for P_α for all $\alpha > \|\overline{y}\|_0$. The proof is similar to those of Eremin and Luenberger. Charalambous' second result is the instance of an exact penalization theorem employing Pennisi's [57] second-order sufficiency conditions. He shows that if the second-order sufficiency condition of Theorem 4.7 is satisfied, then $\overline{x}$ is a local minimum for P_α for all $\alpha > \|\overline{y}\|_0$.

Dolecki and Rolewicz [20] present the deepest first-order results for exact penalization currently available in the literature. They consider a model problem that is somewhat more general than the problem $\mathcal{P}$ and obtain exact penalty results based on a more general notion of subdifferential. In this context, they obtain one of the implications in Theorem 2.1 and a version of Corollary 2.3.1. The Dolecki–Rolewicz paper represents the attempt to extend exact penalization techniques to the nondifferentiable case in infinite-dimensions.

Perhaps the most widely referenced paper on exact penalization is by Han and Mangasarian [33]. Their paper is the most comprehensive and comprehensible study of the subject available in the literature. Han and Mangasarian consider the case in which $C = \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$, $S = \mathbb{R}^n$, and f and g are continuously differentiable. One of the most significant contributions of their paper is the relaxation of the first-order conditions under which an exact penalty parameter for $\mathcal{P}$ exists. Specifically, they show that if the Mangasarian–Fromovitz constraint qualification is satisfied at a strict local solution $\overline{x}$ to $\mathcal{P}$, then there exists an $\overline{\alpha} \geq 0$ such that $\overline{x}$ is a local solution to P_α for all $\alpha \geq \overline{\alpha}$. They establish this result for an arbitrary norm by appealing to the equivalence of norms in finite dimensions. This result is an instance of Corollary 2.4.1 (however, Corollary 2.4.1 does not require that $\overline{x}$ be a strict local solution). They also provide a second-order result that is similar to that of Charalambous. Moreover, they establish the equivalence of stationarity conditions for $\mathcal{P}$ and the minimization of P_α, as is done in Proposition 6.2. They conclude by again establishing Eremin's result for the case of convex programming. The penalty functions they consider are a generalization of the Eremin–Zangwill penalty functions and are based on the work of Bertsekas.

The work of Lasserre [44] appears soon after that of Han and Mangasarian. He considers the case in which $C := \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$, $S = \mathbb{R}^n$, f and g are continuously differentiable, and $\mathbb{R}^m$ is endowed with a weighted ℓ_1 norm. In this case, he establishes a second-order result similar to that of Charalambous. Moreover, he shows that if the active constraint gradients are linearly independent at a local solution to $\mathcal{P}$, then an exact penalty parameter exists for $\mathcal{P}$. This result is different from the corresponding result of Han and Mangasarian, and Pietrzykowski since Lasserre does not assume that the solution is a strict local minimum. Nonetheless, it appears that both of these results are subsumed in the work of Dolecki and Rolewicz. Lasserre also recaptures and extends the result of Luenberger by recognizing the relationship between saddle points of the Lagrangian and local minimum of the exact penalty function.

Fletcher [29, §14.3] considers the same situation as Lasserre. Under the hypothesis that the active constraint gradients are linearly independent, Fletcher [29] is the to recognize the actual equivalence of the first- and second-order optimality conditions for $\mathcal{P}$ and the exact penalty function P_α. Consequently, Fletcher's work is a direct precursor of the results presented in this paper.

Bazaraa and Goode [3] consider the case where $C = \mathbb{R}^m_-$, S is closed, and f_0 and g are continuously differentiable. They establish some extensions to Howe's result using some of the modern techniques of nonsmooth analysis. Moreover, by assuming that S is compact, they obtain global versions of Howe's theorem and give estimates for the value of an exact penalty parameter that are reminiscent of those established by Zangwill.

In [15] Clarke establishes his elementary exact penalization result for the case in which the inclusion constraint $g(x) \in C$ is absent. This result is one of the corner stones of §2 and appears as Theorem 2.3. Clarke's proof should be reviewed by every student of this subject. It is very elementary, requiring only seven short sentences. Clarke also shows that calmness implies the existence of an exact penalty parameter for $\mathcal{P}$ when $C := \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$.

In [59], Polak, Mayne, and Wardi consider the case where $C = \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$, $S = \mathbb{R}^n$, f and g_i, $i = 1, \cdots, s$ are locally Lipschitz, and g_i, $i = s+1, \cdots, m$ are continuously differentiable. In this setting, they establish the equivalence of the stationarity conditions for $\mathcal{P}$ and the minimization of P_α for all α sufficiently large. This result is generalized in Proposition 2.6.

Rosenberg [71] considers the case in which $C = \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$, $S = \mathbb{R}^n$, and f and g are locally Lipschitz functions. He begins by providing local and strict local versions of Clarke's result that calmness implies the existence of an exact penalty parameter. He then reviews the convex programming case and establishes the version of Theorem 4.8, upon which our treatment is based. Rosenberg concludes his study by extending Howe's result to the Lipschitzian case where he provides results that are substantially more general than those of Bazaraa and Goode. For problems of this type he also provides a sharp lower bound for the value of an exact penalty parameter.

Garcia-Palomares [31] examines the case in which $C = \mathbb{R}^s_- \times \{0\}_{\mathbb{R}^{m-s}}$ $S = \mathbb{R}^n$, f and g are continuously differentiable, and $\mathbb{R}^m$ is endowed with the ℓ_∞ norm. The perspective in this paper is quite similar to the one we have taken. His goal is to establish the equivalence between the first- and second-order optimality conditions for $\mathcal{P}$ and P_α. In this regard, he provides versions of some of the results presented in the latter half of §2 and §3. His approach allows a great deal of further insight in the case of the ℓ_∞-norm.

In [50] Mangasarian considers the convex programming case with $C = \mathbb{R}^s_-$, $S = \mathbb{R}^n$, and f and g continuously differentiable. He extends the analysis of Zangwill to provide lower bounds for the value of an exact penalty function under weaker hypotheses.

Recently Conn and Gould [18, 1987] have generalized the ℓ_1 exact penalty function to obtain an exact penalty function for a class of semi-infinite programming problems. They consider both the convex and nonconvex cases, and their results are not covered by those presented in this paper. These new exact penalty functions for semi-infinite programming are quite interesting and deserve much further study.

In [64] Rockafellar studies the case in which $C \subset \mathbb{R}^m$ is the product of intervals, $X \subset \mathbb{R}^n$ is polyhedral convex, and $f_0(x) := \max\{f_{0_j}(x) : j = 1, \cdots, s\}$ where $f_{0_j}, j = 1, \cdots, s$ and g are all continuously differentiable. As in our study, Rockafellar derives the equivalence of first- and second-order optimality conditions for $\mathcal{P}$ and P_α via similar results for convex composite optimization. However, Rockafellar's results rely on the piecewise linear–quadratic case, the theory that he develops in [63].

We conclude by offering our apologies to the many authors we have not mentioned, especially to those who have made significant contributions in the domain of algorithmic development.

A. Appendix.

We proceed to establish Theorem 2.5. For this purpose we will need the following lemmas from Burke and Poliquin [11].

LEMMA A1. *Let $q : \mathbb{R}^n \to \overline{\mathbb{R}}$ be as given in Theorem 2.5. If $\overline{x} \in \operatorname{dom}(q)$ is such that (2.9) holds, then there is a neighborhood U of $\overline{x}$ such that (2.9) is satisfied at every point of $\operatorname{dom}(q) \cap U$.*

Proof. This is a direct consequence of the upper semicontinuity of $\partial f, \partial f^\infty$, ∂g, and $N(\cdot|\operatorname{dom}(h))$. □

LEMMA A2. *Let $h : \mathbb{R}^m \to \overline{\mathbb{R}}$ be as in Theorem 2.5 and let $\{(y_i, z_i)\} \subset$ graph (∂h) be such that $y_i \to y \in \operatorname{dom}(h)$ and $\|z_i\| \uparrow \infty$. Then every cluster point of the sequence $\{z_i/\|z_i\|\}$ is an element of the normal cone to $\operatorname{dom}(h)$ at y.*

LEMMA A3. *Let $h : \mathbb{R}^m \to \overline{\mathbb{R}}$ and $h_\alpha : \mathbb{R}^m \to \mathbb{R}$ be as in Theorem 2.5. If $h_\alpha(\overline{y}) = h(\overline{z}) + \alpha\|\overline{y} - \overline{z}\|$, where $\overline{z} \in \operatorname{dom}(h)$, then $u \in \partial h_\alpha(\overline{y})$ if and only if $u \in \partial h(\overline{z}) \cap (\alpha\mathbb{B}^0)$ and $(\overline{y} - \overline{z}) \in N(u|\alpha\mathbb{B}^0)$.*

The proof of Theorem 2.5 now follows.

Proof. Let ε, $\delta > 0$ be such that $f(x) \geq f(\overline{x})$ for all $x \in \overline{x} + \varepsilon\mathbb{B}$ and (2.9) is satisfied on $\operatorname{dom}(q) \cap (\overline{x} + \delta\mathbb{B})$. Set

$$\xi := 1 + \max\{\|g(x) - g(\overline{x})\| : x \in \overline{x} + \varepsilon\mathbb{B}\},$$

and define

$$\widetilde{h}_\alpha(y) := \inf\{h(z) + \psi(z|g(\overline{x}) + \xi\mathbb{B}) + \alpha\|y - z\| : z \in \mathbb{R}^m\}.$$

Consider the function

$$\widehat{q}_\alpha(x) := \widetilde{q}_\alpha(x) + \varphi(x) + \psi(x|\overline{x} + \varepsilon\mathbb{B}),$$

where $\widetilde{q}_\alpha := f + \widetilde{h}_\alpha \circ g$ and $\varphi(x) := \operatorname{dist}_2^2(x|\overline{x} + \delta\mathbb{B})$. Observe that arg min$\widehat{q}_\alpha$ is nonempty as $\widehat{q}_\alpha$ is lower semicontinuous and $\overline{x} + \varepsilon\mathbb{B}$ is compact. Hence, there is a sequence $\alpha_i \uparrow \infty$ for which there is a corresponding sequence $\{x_i\} \subset \overline{x} + \varepsilon\mathbb{B}$ converging to some element $\widehat{x}$ of $\overline{x} + \varepsilon\mathbb{B}$ such that

$$x_i \in \arg\min \widehat{q}_{\alpha_i}$$

for each $i = 1, 2, \cdots$. Also, from the lower semicontinuity of h and the compactness of $g(\overline{x}) + \xi\mathbb{B}$, there exists for each $i = 1, 2, \cdots$ a y_i in $\mathrm{dom}(h) \cap (g(\overline{x}) + \xi\mathbb{B})$ such that

$$\widetilde{h}_{\alpha_i}(x_i) = h(y_i) + \alpha_i \|y_i - g(x_i)\|.$$

Clearly,

$$q(\overline{x}) \geq \widehat{q}_{\alpha_i}(x_i) \geq \widetilde{q}_{\alpha_i}(x_i). \tag{A.1}$$

Therefore, as $\alpha_i \uparrow \infty$ we have $\|y_i - g(x_i)\| \to 0$ so that $y_i \to g(\widehat{x})$, and thus eventually $y_i \in \mathrm{int}(g(\overline{x}) + \xi\mathbb{B})$, which implies that $\widetilde{q}_{\alpha_i}(x_i) = q_{\alpha_i}(x_i)$. From (A.1) we also obtain that $g(\widehat{x}) \in \mathrm{dom}(h) \cap (g(\overline{x}) + \xi B)$, $\widehat{x} \in \overline{x} + \varepsilon\mathbb{B}$, and

$$q(\overline{x}) \geq q(\widehat{x}) + \varphi(\widehat{x}) \geq q(\widehat{x}).$$

But since $\widehat{x} \in \overline{x} + \varepsilon\mathbb{B}$, the hypotheses imply that $q(\overline{x}) = q(\widehat{x})$ and $\widehat{x} \in \overline{x} + \delta\mathbb{B}$.

We now show that eventually $g(x_i) \in \mathrm{dom}(h)$. Since $x_i \in \arg\min \widehat{q}_{\alpha_i}$ and $x_i \to \widehat{x} \in \overline{x} + \delta\mathbb{B}$, we know that eventually

$$0 \in \partial\widetilde{q}_{\alpha_i}(x_i) + \nabla\varphi(x_i).$$

Hence, by Rockafellar [66, Cor. 5.2.3] and Lemma A.3, eventually there exist $v_i \in \partial f(x_i)$ and $w_i \in \partial\widetilde{h}_{\alpha_i}(g(x_i))$ with

$$w_i \in \partial h(y_i) \text{ and } (g(x_i) - y_i) \in N(w_i | \alpha_i \mathbb{B}^0)$$

(since $N(y_i | g(\overline{x}) + \xi\mathbb{B}) = \{0\}$ as eventually $y_i \in \mathrm{int}(F(\overline{x}) + \xi\mathbb{B})$) such that

$$0 \in v_i + \partial g(x_i)^T w_i + \nabla\varphi(x_i). \tag{A.2}$$

If the sequence $\{(v_i, w_i)\}$ possesses a divergent subsequence $\{(v_i, w_i)\}_J$, then, by Lemma A.2, the sequence $\{(v_i, w_i)/\|(v_i, w_i)\|\}_J$ possesses a cluster point $(\overline{v}, \overline{w})$ with $\overline{v} \in \partial^\infty f(\widehat{x})$, $\overline{w} \in N(g(\widehat{x}) | \mathrm{dom}(h))$, and $\|(\overline{v}, \overline{w})\| = 1$. But for such a cluster point $(\overline{v}, \overline{w})$ we obtain from (3.5) that $0 \in \partial^\infty f(\widehat{x}) + \partial g(\widehat{x})^T w$ which contradicts the choice of δ. Thus the sequence $\{(v_i, w_i)\}$ is bounded. Hence for $\overline{\alpha}$ sufficiently large $\{w_i\} \subset \overline{\alpha}\mathbb{B}^0$ so that $N(w_i | \alpha_i \mathbb{B}^0) = \{0\}$ for all i such that $\alpha_i > \overline{\alpha}$. But then $y_i = g(x_i)$ so that $g(x_i) \in \mathrm{dom}(h)$ whenever $\alpha_i > \overline{\alpha}$. Therefore, for all $\alpha_i > \overline{\alpha}$,

$$q(\overline{x}) \geq \widehat{q}_{\alpha_i}(x_i) \geq \widetilde{q}_{\alpha_i}(x_i) = q_{\alpha_i}(x_i) = q(x_i) \geq q(\overline{x}),$$

so that $\overline{x} \in \arg\min \widehat{q}_{\alpha_i}$. Consequently, $\overline{x}$ is also a local minimizer of q_{α_i} for all $\alpha_i > \overline{\alpha}$. □

Remark. The method of proof also shows that if $\overline{x}$ is a strict local minimizer of q, then it is also a strict local minimizer of q_α.

REFERENCES

[1] J. P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, John Wiley, New York, 1984.

[2] J. W. BANDLER AND C. CHARALAMBOUS, *Nonlinear programming using minimax techniques*, J. Optim. Theory Appl., 13 (1974), pp. 607–619.

[3] J. S. BAZARAA AND J. J. GOODE, *Sufficient conditions for a globally exact penalty function without convexity*, Math. Programming Stud., 19 (1982), pp. 1–15.

[4] A. Ben-Israel, A. Ben-Tal, and S. Zlobec, *Optimality in Non-linear Programming, A Feasible Directions Approach*, Wiley–Intersci. Publ., John Wiley, New York, 1981.

[5] D. P. Bertsekas, *Variable metric methods for constrained optimization using differentiable exact penalty functions*, Proc. 18th Annual Allerton Conference on Communication, Control, and Computing, 1980, pp. 584–593.

[6] ———, *Necessary and sufficient conditions for a penalty function to be exact*, Math. Programming, 9 (1975), pp. 87–89.

[7] J. M. Borwein, *Stability and regular points of inequality systems*, J. Optim. Theory Appl., 48 (1986), pp. 9–52.

[8] J. V. Burke, *Calmness and exact penalization*, SIAM J. Control Optim., 1991.

[9] ———, *An exact penalization viewpoint of constrained optimization*, Technical Report ANL/MCS–TM–95, Mathematics and Computer Science Division, Argonne National Laboratories, Argonne, IL, 1987.

[10] ———, *Second order necessary and sufficient conditions for convex composite N.D.O.*, Math. Programming, 38 (1987), pp. 287–302.

[11] J. V. Burke and R. A. Poliquin, *Optimality conditions for nonfinite valued convex composite functions*, Math. Programming, 1991, to appear.

[12] R. M. Chamberlain, C. Lemarechal, H. C. Pederson, and M. J. D. Powel, *The watchdog technique for forcing convergence in algorithms for constrained optimization*, Math. Programming Stud., 16 (1982), pp. 1–17.

[13] C. Charalambous, *A lower bound for the controlling parameter of the exact penalty function*, Math. Programming, 15 (1978), pp. 278–290.

[14] A. Charnes, W. W. Cooper, and A. Henderson, *An Introduction to Linear Programming*, John Wiley, New York, 1953.

[15] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Canad. Math. Soc. Ser. Monographs Adv. Texts, John Wiley, New York, 1983.

[16] ———, *A new approach to lagrange multipliers*, Math. Oper. Res., 1 (1976), pp. 165–174.

[17] ———, *Generalization gradients and applications*, Trans. Amer. Math. Soc., 205 (1975), pp. 247–262.

[18] A. R. Conn and N. I. M. Gould, *An exact penalty function for semi-infinite programming*, Math. Programming, 37 (1987), pp. 19–40.

[19] A. R. Conn and T. Pietrzykowski, *A penalty function method converging directly to a constrained optimum*, SIAM J. Numer. Anal., 14 (1977), pp. 348– 374.

[20] S. Dolecki and S. Rolewicz, *Exact penalties for local minima*, SIAM J. Control Optim. 17 (1979), pp. 596–606.

[21] M. Edelstein, *A note on nearest points*, Quart. J. Math. Oxford, 21 (1970), pp. 403–405.

[22] I. Ekeland, *On the variational principle*, J. Math. Anal. Appl., 47 (1974), pp. 324–353.

[23] I. I. Eremin, *The penalty method in convex programming*, Soviet Math. Dokl., 8 (1966), pp. 459–462.

[24] J. P. Evans, F. J. Gould, and J. W. Tolle, *Exact penalty functions in nonlinear programming*, Math. Programming, 4 (1973), pp. 72–97.

[25] A. V. Fiacco and G. P. McCormick, *Asymptotic conditions for constrained minimization*, RAC–TP–340, Research Analysis Corporation, McLean, VA, 1968.

[26] R. Fletcher, *Practical Methods of Optimization*, second edition, John Wiley, New York, 1987.

[27] ———, *Penalty functions,* in Mathematical Programming: The State of the Art, Bonn 1982, A. Bachem, M. Grötschel, and B. Korte, eds., Springer–Verlag, Berlin, 1983, pp. 87–114.

[28] ———, *A model algorithm for composite nondifferentiable optimization problems*, Math. Programming Study, 17 (1982), pp. 67–76.

[29] ———, *Practical Methods of Optimization, Volume* 2; *Constrained Optimization*, John Wiley, New York, 1981.

[30] ———, *An exact penalty function for nonlinear programming with inequalities*, Math. Programming, 5 (1973), pp. 129–150.

[31] U. M. Garciá–Palomares, *Connections Among Nonlinear Programming, Minimax, and Exact Penalty Functions*, Computer Science Division, Argonne National Laboratories, Technical Memorandum No. 20, 1983.

[32] J. Gauvin, *A necessary and sufficient regularity condition to have bounded multipliers in nonconvex programming*, Math. Programming, 12 (1977), pp. 136–138.

[33] S.–P. Han and O. L. Mangasarian, *Exact penalty functions in nonlinear programming*, Math. Programming, 17 (1979), pp. 251–269.

[34] ———, *A dual differentiable exact penalty function*, Math. Programming, 14 (1978), pp. 73–86.

[35] M. R. HESTENES, *Optimization Theory: The Finite Dimensional Case*, Robert E. Krieger, New York, reprint, 1981.

[36] ———, *Calculus of Variations and Optimal Control Theory*, John Wiley, New York, 1966.

[37] S. HOWE, *New conditions for exactness of simple penalty functions*, SIAM J. Control Optim., 11 (1973), pp. 378–381.

[38] L. HURWICZ, *Programming in linear spaces*, in Studies in Linear and Non-Linear Programming, K. J. Arrow, L. Hurwicz, and H. Uzawa, eds., Stanford University Press, Stanford, CA, 1958.

[39] A.D. IOFFE, *Necessary and sufficient conditions for a local minimum. 1: A reduction theorem and order conditions*, SIAM J. Control Optim. 17 (1979), pp. 245–250.

[40] ———, *Necessary and sufficient conditions for a local minimum. 2: Conditions of Levitin-Miljutin-Osmolovskii Type*, SIAM J. Control Optim. 17 (1979), pp. 251–265.

[41] ———, *Necessary and sufficient conditions for a local minimum. 3: Second order conditions and augmented duality*, SIAM J. Control Optim. 17 (1979), pp. 266–288.

[42] F. JOHN, *Extremum Problems with Inequalities as Subsidiary Conditions*, Studies and Essays Presented to R. Courant on his 60th birthday, Wiley-Interscience, New York, 1948, pp. 187–204.

[43] H. W. KUHN AND A. W. TUCKER, *Nonlinear Programming*, Proc. 2nd Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, Berkeley, 1961, pp. 481–492.

[44] J. B. LASSERRE, *Exact penalty functions and Lagrange multipliers*, (R.A.I.R.). Automat./Syst. Anal. Control, 14 (1980), pp. 117–125.

[45] E. S. LEVITIN, A. A. MILJUTIN, AND N. P. OSMOLOVSKII, *On conditions for a local minimum in a problem with constraints*, in Mathematical Economics and Functional Analysis, B. S. Mitjagin, ed., Nauka, Moscow, 1974, pp. 139–202 (In Russian).

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

[47] ———, *Control problems with kinks*, IEEE Trans. Automat. Control, 15 (1970), pp. 570–575.

[48] L. A. LYUSTERNIK, *On constrained extrema of functionals*, Matem. Sbornik, 41 (1934), pp. 390–401. (In Russian).

[49] J. MAGUREGUI, *Regular Multivalued Functions and Algorithmic Applications*, Ph.D. thesis, University of Wisconsin at Madison, Madison, WI, 1977.

[50] O. L. MANGASARIAN, *Sufficiency of exact penalty minimization*, SIAM J. Control Optim., 23 (1985), pp. 30–37.

[51] ———, *Nonlinear Programming*, Robert E. Kreiger, New York, reprint, 1979.

[52] O. L. MANGASARIAN AND S. FROMOVITZ, *The Fritz John necessary optimality conditions in the presence of equality and inequality constraints*, J. Math. Anal. Appl.,17 (1967), pp. 37–47.

[53] P. MICHEL AND J.-P. PENOT, *Calcul sous-differential pour des fonctions lipschitziennes et non lipschitziennes*, C. R. Acad. Sci. Paris, 298 (1984), pp. 269–272.

[54] B. S. MORDUKHOVICH, *Approximation Methods in Optimization and Control Problems*, Main Physical and Mathematical Editions, 1988, Moscow, Nauka. (In Russian).

[55] D. D. MORRISON, *Optimization by least squares*, SIAM J. Numer. Anal., 5 (1968), pp. 83–88.

[56] V. H. NGUYEN, J.-J. STRODIOT, AND R. MIFFLIN, *On Conditions to Have Bounded Multipliers in Locally Lipschitz Programming*, Math. Programming, 18 (1980), pp. 100–106.

[57] L. PENNISI, *An indirect proof for the problem of Lagrange with differential inequalities as added side conditions*, Trans. Amer. Math. Soc., 74 (1953), pp. 177–198.

[58] T. PIETRZYKOWSKI, *An exact potential method for constrained maxima*, SIAM J. Numer. Anal., 6 (1969), pp. 299–304.

[59] E. POLAK, D. Q. MAYNE, AND Y. WARDI, *On the extension of constrained optimization algorithms from differentiable to non-differentiable problems*, SIAM J. Control Optim., 21 (1983), pp. 179–203.

[60] G. DI PILLO AND L. GRIPPO, *A continuously differentiable exact penalty function for nonlinear programming problems with inequality constraints*, SIAM J. Control Optim., 23 (1985), pp. 72–84.

[61] S. M. ROBINSON, *Stability theory for systems of inequalities. Part* II. *Differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497–513.

[62] ———, *Regularity and stability for convex multivalued functions*, Math. Oper. Res., 1 (1976), pp. 130–143.

[63] R. T. ROCKAFELLAR, *Second-order optimality conditions in nonlinear programming obtained by way of epi-derivatives*, Math. Oper. Res., to appear.

[64] ———, *-and second-order epi-differentiability in nonlinear programming*, Trans. Amer. Math. Soc., 307 (1988), pp. 75–108.

[65] ———, *Proximal subgradients, marginal values, and augmented Lagrangians in nonconvex optimization*, Math. Oper. Res., 6 (1981), pp. 424–436.

[66] ———, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Anal. Theory Meth. Appl., 9 (1985), pp. 665–698.

[67] ———, *Lipschitzian properties of multifunctions*, Nonlinear Anal. TMA, 9 (1985), pp. 867–885.

[68] ———, *Directionally Lipschitzian functions and subdifferential calculus*, Proc. London Math. Soc., 39 (1979), pp. 331–355.

[69] ———, *Augmented Lagrange multiplier functions and duality in nonconvex programming*, SIAM J. Control Optim., 12 (1974), pp. 268–285.

[70] ———, *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[71] E. ROSENBERG, *Exact penalty functions and stability in locally Lipschitz programming*, Math. Programming, 30 (1984), pp. 340–356.

[72] M. SLATER, *Lagrange multipliers revisited: a contribution to non-linear programming*, Cowles Commission Discussion Paper, Math., 403, 1950.

[73] C. URSESCU, *Multifunctions with closed convex graph*, Czechoslavak Math. J., 25 (1975), pp. 438–441.

[74] L. P. VLASOV, *Approximative properties of sets in normed linear spaces*, Russian Math. Surveys, 28 (1975), pp. 1–61.

[75] W. I. ZANGWILL, *Nonlinear programming via penalty functions*, Management Sci., 13 (1967), pp. 344–358.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 5, pp. 999-1018, September 1991

001

RECURSIVE STOCHASTIC ALGORITHMS FOR GLOBAL OPTIMIZATION IN $\mathbb{R}^d$*

SAUL B. GELFAND† AND SANJOY K. MITTER‡

Abstract. An algorithm of the form $X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k$, where $U(\cdot)$ is a smooth function on $\mathbb{R}^d$, $\{\xi_k\}$ is a sequence of $\mathbb{R}^d$-valued random variables, $\{W_k\}$ is a sequence of independent standard d-dimensional Gaussian random variables, $a_k = A/k$ and $b_k = \sqrt{B}/\sqrt{k \log\log k}$ for k large, is considered. An algorithm of this type arises by adding slowly decreasing white Gaussian noise to a stochastic gradient algorithm. It is shown, under suitable conditions on $U(\cdot)$, $\{\xi_k\}$, A, and B, that X_k converges in probability to the set of global minima of $U(\cdot)$. No prior information is assumed as to what bounded region contains a global minimum. The analysis is based on the asymptotic behavior of the related diffusion process $dY(t) = -\nabla U(Y(t))dt + c(t)dW(t)$, where $W(\cdot)$ is a standard d-dimensional Wiener process and $c(t) = \sqrt{C}/\sqrt{\log t}$ for t large.

Key words. global optimization, random optimization, simulated annealing, stochastic gradient algorithms, diffusions

AMS(MOS) subject classifications. 65K10, 90C30, 60J60

1. Introduction. In this paper we consider a class of algorithms for finding a global minimum of a smooth function $U(x)$, $x \in \mathbb{R}^d$. Specifically, we analyze the convergence of a modified stochastic gradient algorithm

$$X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k, \tag{1.1}$$

where $\{\xi_k\}$ is a sequence of $\mathbb{R}^d$-valued random variables, $\{W_k\}$ is a sequence of standard d-dimensional independent Gaussian random variables, and $\{a_k\}$, $\{b_k\}$ are sequences of positive numbers with $a_k, b_k \to 0$. An algorithm of this type arises by artificially adding the $b_k W_k$ term (via a Monte Carlo simulation) to a standard stochastic gradient algorithm,

$$Z_{k+1} = Z_k - a_k(\nabla U(Z_k) + \xi_k). \tag{1.2}$$

Algorithms like (1.2) arise in a variety of optimization problems including adaptive filtering, identification, and control; here the sequence $\{\xi_k\}$ is due to noisy or imprecise measurements of $\nabla U(\cdot)$ (cf. [1]). The asymptotic behavior of $\{Z_k\}$ has been extensively studied. Let S and S^* be the set of local and global minima of $U(\cdot)$, respectively. It can be shown, for example, that if $U(\cdot)$ and $\{\xi_k\}$ are suitably behaved, $a_k = A/k$ for k large, and $\{Z_k\}$ is bounded, then $Z_k \to S$ as $k \to \infty$ with probability one. However, in general, $Z_k \not\to S^*$ (unless of course $S = S^*$). The idea behind the additional $b_k W_k$ term in (1.1) compared with (1.2) is that if b_k tends to zero slowly enough, then possibly $\{X_k\}$ (unlike $\{Z_k\}$) will avoid getting trapped in a strictly local minimum of $U(\cdot)$. In fact, we will show that if $U(\cdot)$ and $\{\xi_k\}$ are suitably behaved, $a_k = A/k$ and $b_k^2 = B/k \log\log k$ for k large with $B/A > C_0$ (where C_0 is a positive constant that depends only on $U(\cdot)$), and $\{X_k\}$ is tight, then $X_k \to S^*$ as $k \to \infty$ in probability. We also give a condition for the tightness of $\{X_k\}$. We remark that in [1] both probability one and

* Received by the editors January 10, 1990; accepted for publication (in revised form) July 26, 1990. This research was supported by National Science Foundation grant ECS-8910073, Air Force Office of Scientific Research contract 89-0276B, and Army Research Office contract DAAL03-86-K-0171 (Center for Intelligent Control Systems).

† School of Electrical Engineering, Purdue University, West Lafayette, Indiana 47907.

‡ Department of Electrical Engineering and Computer Science, and Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

weak convergence of $\{Z_k\}$ are treated. Furthermore, convergence of Z_k to S is established under very weak conditions on $\{\xi_k\}$ assuming that $\{Z_k\}$ is bounded. Here the convergence of X_k to S^* is established under somewhat stronger conditions on $\{\xi_k\}$ assuming that $\{X_k\}$ is tight (which is weaker than boundedness).

An algorithm like (1.1) was first proposed and analyzed by Kushner [2]. However, the analysis in [2] required that the trajectories of $\{X_k\}$ lie within a fixed ball (which was achieved by modifying (1.1) near the boundary of the ball). Hence, the version of (1.1) in [2] is only suitable for optimizing $U(\cdot)$ over a compact set. Some other differences between the results presented here and in [2] include conditions on $\{a_k\}$, $\{b_k\}$, and $\{\xi_k\}$, and also the method of analysis; these are discussed below.

The analysis of the convergence of $\{Z_k\}$ is usually based on the asymptotic behavior of the *associated ordinary differential equation* (ODE)

$$\dot{z}(t) = -\nabla U(z(t)) \tag{1.3}$$

(cf. [1], [3]). This motivates our analysis of the convergence of $\{X_k\}$ based on the asymptotic behavior of the *associated stochastic differential equation* (SDE)

$$dY(t) = -\nabla U(Y(t))\,dt + c(t)\,dW(t), \tag{1.4}$$

where $W(\cdot)$ is a standard d-dimensional Wiener process and $c(\cdot)$ is a positive function with $c(t)\to 0$ as $t\to\infty$. The diffusion $Y(\cdot)$ has been called continuous simulated annealing. In this context, $U(x)$ is called the energy of state x and $T(t) = c^2(t)/2$ is called the temperature at time t. Continuous simulated annealing was first suggested in [4] and [5] for global optimization problems that arise in image processing applications with continuous grey levels. Now the asymptotic behavior of $Y(t)$ as $t\to\infty$ has been studied intensively by a number of researchers. In [2] and [5], convergence results were obtained by considering a version of (1.4) with a reflecting boundary; in [6] and [7] the reflecting boundary was removed. Our analysis of $\{X_k\}$ is based on the analysis of $Y(\cdot)$ developed by Chiang, Hwang, and Sheu [7] who prove the following result: if $U(\cdot)$ is well behaved and $c^2(t) = C/\log t$ for t large with $C > C_0$ then $Y(t)\to S^*$ as $t\to\infty$ in probability. The main difficulty associated with using $Y(\cdot)$ to analyze $\{X_k\}$ is that we must deal with long time intervals and slowly decreasing (unbounded) Gaussian noise.

We make some further remarks on the differences between the results and methods in this paper as compared with [2]. We first note that in [2] the modified version of (1.1), which constrains the trajectories of $\{X_k\}$ to lie within a fixed ball, is analyzed for $a_k = b_k = A/\log k$, k large. Although a detailed asymptotic description of $\{X_k\}$ is obtained for this case, in general, $X_k \not\to S^*$ unless $\xi_k = 0$. The reason for this is intuitively clear: even if $\{\xi_k\}$ is bounded, $a_k\xi_k$ and $a_k W_k$ can be of the same order, and hence can interfere with each other. On the other hand, here we allow $\{\xi_k\}$ with unbounded variance, in particular, $E\{|\xi_k|^2\} = O(k^\gamma)$ and $\gamma < 1$. This has important implications when $\nabla U(\cdot)$ is not measured exactly. We also note that the analysis in [2] is different from that done here, in that in [2] the behavior of $\{X_k\}$ is obtained by deriving various large deviations estimates of Donsker–Varadhan type, whereas here we obtain the behavior of $\{X_k\}$ directly from the corresponding behavior of $Y(\cdot)$. It should be pointed out that in a certain sense the results in [2] are also stronger than those presented here, because the large deviation approach in [2] treats the whole tail of the process $\{X_k\}$, while only "local" type results are discussed here. However, from our point of view the most significant difference between our work and that done in [2] (and more generally in other work on global optimization such as [8]) is that we deal with unbounded processes and establish the convergence of an algorithm that finds a global minimum of a function when it is not known a priori what bounded region contains such a point.

The paper is organized as follows. In § 2 we state our assumptions and main result. In § 3 we take up the proof of this result. In § 4 we prove a general tightness criterion, which is then used in § 5 to establish tightness and ultimately convergence for two example algorithms.

2. Main result. In this section we present our main result on the convergence of the discrete time algorithm

$$X_{k+1}=X_k-a_k(\nabla U(X_k)+\xi_k)+b_kW_k, \qquad k\geqq 0, \tag{2.1}$$

which is closely related to the continuous time algorithm

$$dY(t)=-\nabla U(Y(t))\,dt+c(t)\,dW(t), \qquad t\geqq 0. \tag{2.2}$$

Here $U(\cdot)$ is a smooth function on $\mathbb{R}^d$, $\{\xi_k\}$ is a sequence of $\mathbb{R}^d$-valued random variables, $\{W_k\}$ is a sequence of independent d-dimensional Gaussian random variables with $E\{W_k\}=0$ and $E\{W_k\otimes W_k\}=I$ (the identity matrix), $W(\cdot)$ is a standard d-dimensional Wiener process, and

$$a_k=\frac{A}{k},\quad b_k^2=\frac{B}{k\log\log k},\quad k \text{ large},$$

$$c^2(t)=\frac{C}{\log t},\quad t \text{ large},$$

where A, B, and C are positive constants with $C=B/A$. Further conditions on $U(\cdot)$, $\{\xi_k\}$, and $\{W_k\}$ will be discussed below. It will be useful to define a continuous-time interpolation of $\{X_k\}$. Let

$$t_k=\sum_{n=0}^{k-1}a_n, \qquad k\geqq 0,$$

and

$$X(t)=X_k, \quad t\in[t_k,t_{k+1}), \quad k\geqq 0.$$

In the sequel we assume some or all of the following conditions (α and β are constants whose values will be specified later):

(A1) $U(\cdot)$ is a C^2 function from $\mathbb{R}^d$ to $[0,\infty)$ such that

$$\min U(x)=0,$$

$$U(x)\to\infty \quad\text{and}\quad |\nabla U(x)|\to\infty \quad \text{as } |x|\to\infty,$$

$$\inf\,(|\nabla U(x)|^2-\Delta U(x))>-\infty.$$

(A2) For $\varepsilon>0$ let

$$d\pi^\varepsilon(x)=\frac{1}{Z^\varepsilon}\exp\left(-\frac{2U(x)}{\varepsilon^2}\right)dx, \qquad Z^\varepsilon=\int\exp\left(-\frac{2U(x)}{\varepsilon^2}\right)dx<\infty.$$

π^ε has a unique weak limit π as $\varepsilon\to 0$.

(A3) $$\varliminf_{|x|\to\infty}\left\langle\frac{\nabla U(x)}{|\nabla U(x)|},\frac{x}{|x|}\right\rangle>L(d), \qquad L(d)=\left(\frac{4d-4}{4d-3}\right)^{1/2}.$$

(A4) For $k=0,1,\cdots$ let $\mathscr{F}_k=\sigma(X_0, W_0,\cdots, W_{k-1}, \xi_0,\cdots,\xi_{k-1})$. Let K be a compact subset of $\mathbb{R}^d$. There exists $L>0$ such that

$$E\{|\xi_k|^2|\mathscr{F}_k\}\leqq La_k^\alpha, \quad |E\{\xi_k|\mathscr{F}_k\}|\leqq La_k^\beta, \quad \forall X_k\in K, \quad \text{w.p.1.}$$

W_k is independent of $\mathscr{F}_k$.

We note that the measure π concentrates on S^*, the global minima of $U(\cdot)$. The existence of π and a simple characterization in terms of the Hessian of $U(\cdot)$ is discussed in [9]. In [7], (A1) and (A2) were needed for the analysis of $Y(t)$ as $t\to\infty$; here we also need (A3) and (A4) for the analysis of X_k as $k\to\infty$. Assumption (A3) asserts that $\nabla U(x)$ has a sufficiently large radial component for $|x|$ large. This condition will be used to extend an escape time estimate for $\{X_k\}$ from a bounded region in the $d=1$ case to the $d>1$ case (see Lemma 4). It may be possible to replace $L(d)$ by 0 in (A3) but we have not been able to do so (except of course for $d=1$). Note that (A3) is implied by (A1) when $d=1$.

For a process $Z(\cdot)$ and a function $f(\cdot)$, let $E_{t_1,z_1}\{f(Z(t))\}$ denote conditional expectation given $Z(t_1)=z_1$, and let $E_{t_1,z_1;t_2,z_2}\{f(Z(t))\}$ denote conditional expectation given $Z(t_1)=z_1$ and $Z(t_2)=z_2$ (more precisely, these are suitable fixed versions of conditional expectations). Also for a measure $\mu(\cdot)$ and a function $f(\cdot)$ let $\mu(f)=\int f\,d\mu$.

In [7] it was shown that there exists a constant C_0 (denoted there by c_0) that plays a critical role in the convergence of $Y(t)$ as $t\to\infty$. C_0 has an interpretation in terms of the action functional [10] for the perturbed dynamical systems

$$dY^\varepsilon(t)=-\nabla U(Y^\varepsilon(t))\,dt+\varepsilon\,dW(t). \tag{2.3}$$

Now for $\phi(\cdot)$ an absolutely continuous function on $\mathbb{R}^d$, the (normalized) action functional for (2.3) is given by

$$I(t,x,y)=\inf_{\substack{\phi(0)=x\\ \phi(t)=y}}\frac{1}{2}\int_0^t|\dot\phi(s)+\nabla U(\phi(s))|^2\,ds.$$

According to [7]

$$C_0=\tfrac{3}{2}\sup_{x,y\in S_0}(V(x,y)-2U(y)),$$

where $V(x,y)=\lim_{t\to\infty}I(t,x,y)$ and S_0 is the set of all the stationary points of $U(\cdot)$, i.e., $S_0=\{x\colon \nabla U(x)=0\}$; see [7] for a further discussion of C_0 including some examples. Here is the Chiang–Hwang–Sheu theorem on the convergence of $Y(t)$ as $t\to\infty$.

THEOREM 1 [7]. *Assume* (A1) *and* (A2) *hold. Then for $C>C_0$ and any bounded continuous function $f(\cdot)$ on $\mathbb{R}^d$*

$$\lim_{t\to\infty}E_{0,y_0}\{f(Y(t))\}=\pi(f) \tag{2.4}$$

uniformly for y_0 in a compact set.

Let $K_1\subset\mathbb{R}^d$ and let $\{X_k^{x_0}\}$ denote the solution of (2.1) with $X_0=x_0$. We say that $\{X_k^{x_0}\colon k\geqq 0, x_0\in K_1\}$ is tight if given $\varepsilon>0$ there exists a compact $K_2\subset\mathbb{R}^d$ such that $P_{0,x_0}\{X_k\in K_2\}>1-\varepsilon$ for all $k\geqq 0$ and $x_0\in K_1$. Here is our theorem on the convergence of X_k as $k\to\infty$.

THEOREM 2. *Assume* (A1)-(A4) *hold with $\alpha>-1$ and $\beta>0$. Also assume that $\{X_k^{x_0}\colon k\geqq 0, x_0\in K\}$ is tight for K a compact set. Then for $B/A>C_0$ and any bounded* *continuous function* $f(\cdot)$ *on* $\mathbb{R}^d$

$$\lim_{k\to\infty} E_{0,x_0}\{f(X_k)\}=\pi(f) \tag{2.5}$$

uniformly for x_0 *in a compact set.*

Remark. We specifically separate the question of tightness from convergence in Theorem 2. It is appropriate to do this because sometimes it is convenient to first prove tightness and then to put an algorithm into the form of (2.1) to prove convergence. In § 4, we actually give a condition for tightness of a class of algorithms somewhat more general than (2.1), and then use it in § 5 to prove tightness and ultimately convergence for two example algorithms.

Since π concentrates on S^*, we have, of course, that (2.4) and (2.5) imply $Y(t)\to S^*$ as $t\to\infty$ and $X_k\to S^*$ as $k\to\infty$ in probability, respectively.

The proof of Theorem 2 requires the following two lemmas. Let $\beta(\cdot)$ be defined by

$$\int_s^{\beta(s)} \frac{\log s}{\log u}\,du=s^{2/3}, \qquad s>1.$$

Note that $s+s^{2/3}\leqq\beta(s)\leqq s+2s^{2/3}$ for s large.

LEMMA 1 [7]. *Assume the conditions of Theorem* 1. *Then for any bounded continuous function* $f(\cdot)$ *on* $\mathbb{R}^d$

$$\lim_{s\to\infty}(E_{s,x}\{f(Y(\beta(s)))\}-\pi^{c(s)}(f))=0$$

uniformly for x *in a compact set.*

LEMMA 2. *Assume the conditions of Theorem* 2. *Then for any bounded continuous function* $f(\cdot)$ *on* $\mathbb{R}^d$

$$\lim_{s\to\infty}(E_{0,x_0;s,x}\{f(X(\beta(s)))\}-E_{s,x}\{f(Y(\beta(s)))\})=0$$

uniformly for x_0 *in a compact set and all* x.

Lemma 1 is proved in Lemmas 1–3 of [7]. Lemma 2 is proved in § 3. Note that these lemmas involve approximation on increasingly large time intervals: $\beta(s)-s\geqq s^{2/3}\to\infty$ as $s\to\infty$. We now show how these lemmas may be combined to prove Theorem 2.

Proof of Theorem 2. Since $\beta(s)$ is continuous and $\beta(s)\to\infty$ as $s\to\infty$, it is enough to show that

$$\lim_{s\to\infty} E_{0,x_0}\{f(X(\beta(s)))\}=\pi(f) \tag{2.6}$$

uniformly for x_0 in a compact set. We have for $r>0$

$$\begin{aligned}
&|E_{0,x_0}\{f(X(\beta(s)))\}-\pi(f)|\\
&\leqq\int P_{0,x_0}\{X(s)\in dx\}|E_{0,x_0;s,x}\{f(X(\beta(s)))\}-\pi(f)| \qquad (2.7)\\
&\leqq\int_{|x|\leqq r} P_{0,x_0}\{X(s)\in dx\}|E_{0,x_0;s,x}\{f(X(\beta(s)))\}-\pi(f)|+2\|f\|P_{0,x_0}\{|X(s)|>r\}.
\end{aligned}$$

Now by the tightness assumption

$$\sup_{s\geqq0} P_{0,x_0}\{|X(s)|>r\}\to0 \quad \text{as } r\to\infty. \tag{2.8}$$

Also by Lemmas 1 and 2 and assumption (A2)

$$\tag{2.9}\begin{aligned}&\sup_{|x|\leqq r}|E_{0,x_0;s,x}\{f(X(\beta(s)))\}-\pi(f)|\\ &\quad\leqq \sup_{|x|\leqq r}|E_{0,x_0;s,x}\{f(X(\beta(s)))\}-E_{s,x}\{f(Y(\beta(s)))\}|\\ &\qquad+\sup_{|x|\leqq r}|E_{s,x}\{f(Y(\beta(s)))\}-\pi^{c(s)}(f)|\\ &\qquad+|\pi^{c(s)}(f)-\pi(f)|\to 0\quad \text{as } s\to\infty.\end{aligned}$$

Combining (2.7)-(2.9) and letting $s\to\infty$ and then $r\to\infty$ gives (2.6) and hence the theorem. □

3. Proof of Lemma 2. Before proceeding with the proof of Lemma 2 we address the following technical issue. Observe that Lemma 2 is not concerned with the joint probability law of $X(\cdot)$ and $Y(\cdot)$. Hence without loss of generality we can and will assume that

$$W_k=a_k^{-1/2}(W(t_{k+1})-W(t_k)),$$

and that the following assumption holds in place of (A4):

(A4′) For $k=0,1,\cdots$ let $\mathscr{F}_k=\sigma(X_0, Y_0, \xi_0,\cdots,\xi_{k-1}, W(s), 0\leqq s\leqq t_k)$. Let K be a compact subset of $\mathbb{R}^d$. There exists $L>0$ such that

$$E\{|\xi_k|^2|\mathscr{F}_k\}\leqq La_k^\alpha,\quad |E\{\xi_k|\mathscr{F}_k\}|\leqq La_k^\beta\quad \forall X_k\in K,\quad \text{w.p.1.}$$

$W(t)-W(t_k)$ is independent of $\mathscr{F}_k$ for $t>t_k$.

It will also be convenient to define

$$c^2(t_k)=\frac{C}{\log\log k},\quad k \text{ large},$$

and to let $c^2(\cdot)$ be a piecewise linear interpolation of $\{c^2(t_k)\}$. Note that $c^2(t)\sim C/\log t$, and since $C=B/A$ we have $\sqrt{a_k}c(t_k)=b_k$.

In the sequel, $c_1, c_2,\cdots$ denote positive constants whose value may change from proof to proof.

The proof of Lemma 2 is based on the following three lemmas. For s, $R>0$ define the exit times

$$\sigma(s,R)=\inf\{t\geqq s\colon |X(t)|>R\},$$

$$\tau(s,R)=\inf\{t\geqq s\colon |Y(t)|>R\}.$$

LEMMA 3 [7, p. 745]. *Assume the conditions of Theorem* 1. *Then given* $r>0$ *there exists* $R>r$ *such that*

$$\lim_{s\to\infty}P_{s,x}(\tau(s,R)>\beta(s)\}=1$$

uniformly for $|x|\leqq r$.

LEMMA 4. *Assume the conditions of Theorem* 2. *Then given* $r>0$ *there exists* $R>r$ *such that*

$$\lim_{s\to\infty}P_{0,x_0;s,x}\{\sigma(s,R)>\beta(s)\}=1$$

uniformly for $|x|\leqq r$ *and all* x_0.

LEMMA 5. *Assume the conditions of Theorem* 2. *Then for* $0<r<R$

$$\lim_{s\to\infty} E_{0,x_0;s,x}\{|X(\beta(s))-Y(\beta(s))|^2, \sigma(s,R)\wedge\tau(s,R)>\beta(s)\}=0$$

uniformly for $|x|\leqq r$ *and all* x_0.

The proofs of Lemmas 4 and 5 are given below. We now show how these lemmas may be combined to prove Lemma 2.

Proof of Lemma 2. Given $r>0$, choose $R>r$ as in Lemmas 3 and 4. Fix $s>0$ for the moment and let $\sigma=\sigma(s,R)$ and $\tau=\tau(s,R)$. Henceforth assume all quantities are conditioned on $X(0)=x_0$, $X(s)=Y(s)=x$, and $|x|\leqq r$. We have

$$\begin{aligned}&|E(\{f(X(\beta(s)))\}-E\{f(Y(\beta(s)))\}|\\ &\quad\leqq E\{|f(X(\beta(s)))-f(Y(\beta(s)))|, \sigma\wedge\tau>\beta(s)\}+2\|f\|P\{\sigma\wedge\tau\leqq\beta(s)\}.\end{aligned}\tag{3.1}$$

Now by Lemmas 3 and 4

$$P\{\sigma\wedge\tau\leqq\beta(s)\}\to 0 \quad \text{as } s\to\infty.\tag{3.2}$$

Also, since $f(\cdot)$ is uniformly continuous on a compact, given $\varepsilon>0$ there exists $\delta>0$ such that $|f(u)-f(v)|<\varepsilon$ whenever $|u-v|<\delta$ and $|u|,|v|\leqq R$. Hence using the Chebyshev inequality and Lemma 5

$$\begin{aligned}&E\{|f(X(\beta(s)))-f(Y(\beta(s)))|, \sigma\wedge\tau>\beta(s)\}\\ &\quad\leqq 2\|f\|P\{|X(\beta(s))-Y(\beta(s))|\geqq\delta, \sigma\wedge\tau>\beta(s)\}+\varepsilon\\ &\quad\leqq\frac{2\|f\|}{\delta^2}E\{|X(\beta(s))-Y(\beta(s))|^2, \sigma\wedge\tau>\beta(s)\}+\varepsilon\to\varepsilon \quad \text{as } s\to\infty.\end{aligned}\tag{3.3}$$

Combining (3.1)-(3.3) and letting $s\to\infty$ and then $\varepsilon\to 0$ gives the lemma. □

The proofs of Lemmas 4 and 5 involve comparisons between $X(\cdot)$ and $Y(\cdot)$. Define $\zeta(\cdot,\cdot)$ by

$$Y(t)=Y(s)-(t-s)(\nabla U(Y(s))+\zeta(s,t))+c(s)(W(t)-W(s))$$

for $t\geqq s\geqq 0$. To compare $X(\cdot)$ and $Y(\cdot)$ we will need statistics for $\zeta(\cdot,\cdot)$.

PROPOSITION 1. *For every* $R>0$

$$E_{s,y}\{|\zeta(s,t\wedge\tau(s,R))|^2\}=O(|t-s|)$$

as $t\downarrow s$, *uniformly for* $s\geqq 0$ *and all* y.

Proof. In this proof we can and will assume that $\nabla U(\cdot)$ is a bounded and Lipschitz function on $\mathbb{R}^d$ (since $|Y(u)|\leqq R$ for $s\leqq u\leqq t\wedge\tau(s,R)$ we can modify $U(x)$ for $|x|>R$ without loss of generality). Fix $s\geqq 0$ and let $\tau=\tau(s,R)$. Henceforth assume all quantities are conditioned on $Y(s)=y$. Now for $t\geqq s$ we can write

$$(t-s)\zeta(s,t\wedge\tau)=\int_s^{t\wedge\tau}(\nabla U(Y(u))-\nabla U(Y(s)))\,du-\int_s^{t\wedge\tau}(c(u)-c(s))\,dW(u).\tag{3.4}$$

Let d_1 and d_2 be Lipschitz constants for $\nabla U(\cdot)$, $c(\cdot)$, respectively. Under our assumptions on $\nabla U(\cdot)$ and $c(\cdot)$ it is well known (cf. [11]) that $E\{|Y(u)-Y(s)|^2\}=O(|u-s|)$ as $u\downarrow s$, uniformly for $s\geqq 0$ and all y. Hence

$$\begin{aligned}E\left\{\left|\int_s^{t\wedge\tau}(\nabla U(Y(u))-\nabla U(Y(s)))\,du\right|^2\right\}&\leqq d_1^2E\left\{\left(\int_s^t|Y(u)-Y(s)|\,du\right)^2\right\}\\ &\leqq 2d_1^2(t-s)\int_s^t E\{|Y(u)-Y(s)|^2\}\,du\\ &=O((t-s)^3)\end{aligned}\tag{3.5}$$

and

$$\begin{aligned} E\left\{\left|\int_s^{t\wedge\tau} (c(u)-c(s))\,dW(u)\right|^2\right\} &\leqq \int_s^t (c(u)-c(s))^2\,du \\ &\leqq d_2^2\int_s^t (u-s)^2\,du = O((t-s)^3) \end{aligned} \tag{3.6}$$

as $t\downarrow s$, uniformly for $s\geqq 0$ and all y. The proposition follows from (3.4)-(3.6). □

COROLLARY 1. *Given $R>0$, let $\zeta_k=\zeta(t_k, t_{k+1}\wedge\tau(t_k, R))$. Then there exists $M>0$ such that*

$$E\{|\zeta_k|^2|\mathscr{F}_k\}\leqq Ma_k, \quad |E\{\zeta_k|\mathscr{F}_k\}|\leqq Ma_k^{1/2} \quad \text{w.p.1.}$$

Proof. Observe that ζ_k is $\{Y(t_k), W(t)-W(t_k), t_k<t\leqq t_{k+1}\}$ measurable. Since $Y(t_k)$ is $\mathscr{F}_k$ measurable and $\{W(t)-W(t_k), t_k<t\leqq t_{k+1}\}$ is independent of $\mathscr{F}_k$, we must have $P\{\zeta_k\in\cdot\,|\mathscr{F}_k\}=P\{\zeta_k\in\cdot\,|Y(t_k)\}$ w.p.1. The corollary now follows from Proposition 1 and Holder's inequality. □

3.1. Proof of Lemma 4. The idea behind this proof is to compare $X(t)$ and $Y(t)$ in such a way as to eliminate the slowly decreasing Gaussian noise (i.e., the b_kW_k term) between them. Once the decreasing Gaussian noise is eliminated, we can control the deviation of $X(t)$ from $Y(t)$ over increasingly large time intervals and ultimately obtain the escape time estimate for $X(t)$ from a bounded region from that for $Y(t)$ in Lemma 3. It seems very difficult to work directly with the continuous-time interpolation $X(t)$.

For each n let k_n be the integer that satisfies $\beta(t_n)\in[t_{k_n}, t_{k_{n+1}})$. We show there exists $R>r$ such that

$$\lim_{n\to\infty} P_{0,x_0;t_n,x}\{\sigma(t_n, R)>t_{k_n}\}=1 \tag{3.7}$$

uniformly for $|x|\leqq r$ and all x_0. The lemma then follows by some minor details that are omitted.

By Lemma 3 there exists $R_1>r$ such that

$$\lim_{n\to\infty} P_{t_n,x}\{\tau(t_n, R_1)>t_{k_n}\}=1$$

uniformly for $|x|\leqq r$. Hence (3.7) will follow if we can show that there exists $R>r$ such that

$$\lim_{n\to\infty} P_{0,x_0;t_n,x}\{\sigma(t_n, R)\leqq t_{k_n}, \tau(t_n, R_1)>t_{k_n}\}=0 \tag{3.8}$$

uniformly for $|x|\leqq r$ and all x_0. We first assume $d=1$ (the scalar case) and then generalize to $d>1$. The generalization to $d>1$ requires (A3).

Proof for $d=1$. In view of (A1) there exists $R_2>R_1$ such that

$$\sup_{x\leqq -R_2} U'(x)<\inf_{|x|\leqq R_1} U'(x), \qquad \inf_{x\geqq R_2} U'(x)>\sup_{|x|\leqq R_1} U'(x).$$

Let $R_3=R_2+1$ and $R_4=2R_3+3R_1$. We show that (3.8) holds with $R=R_4$.

Fix n for the moment and let $\sigma=\sigma(t_n, R_4)$, $\tau=\tau(t_n, R_1)$. Let

$$\zeta_k=\zeta(t_k, t_{k+1}\wedge\tau(t_k, R_1))$$

and

$$Y_{k+1}=Y_k-a_k(\nabla U(Y_k)+\zeta_k)+b_kW_k.$$

Note that if $Y(t_n)=Y_n$ and $\tau\geqq t_k\geqq t_n$, then $Y(t_k)=Y_k$. Henceforth assume all quantities are conditioned on $X(0)=X_0=x_0$, $X(t_n)=X_n=Y(t_n)=Y_n=x$, $|x|\leqq r$.

We proceed by observing that if the event $\{\sigma\leqq t_{k_n}\}$ occurs then either

- At some time k, $n\leqq k<k_n$, X_k jumps from $[-R_4, R_2]$ to (R_3,∞), or from $[-R_2, R_4]$ to $(-\infty,-R_3)$;
- At some time k, $n\leqq k<k_n$, X_k jumps from $[-R_4, R_2]$ to $(R_2, R_3]$, and exits from $(R_2, R_4]$ to (R_4,∞) at some time l, $k<l\leqq k_n$;
- At some time k, $n\leqq k<k_n$, X_k jumps from $[-R_2, R_4]$ to $[-R_3,-R_2)$, and exits from $[-R_4,-R_2)$ to $(-\infty,-R_4)$ at some time l, $k<l\leqq k_n$.

Now define $\mathscr{F}_k$-stopping times:

$$\begin{aligned}
\mu_1^+&=\inf\{k>n\colon X_{k-1}\leqq R_2,\ R_2<X_k\leqq R_3\},\\
\nu_1^+&=\inf\{k>\mu_1^+\colon X_k\leqq R_2\},\\
\mu_2^+&=\inf\{k>\nu_1^+\colon X_{k-1}\leqq R_2,\ R_2<X_k\leqq R_3\},\\
\nu_2^+&=\inf\{k>\mu_2^+\colon X_k\leqq R_2\},\\
&\ \vdots
\end{aligned}$$

and

$$\begin{aligned}
\mu_1^-&=\inf\{k>n\colon X_{k-1}\geqq -R_2,\ -R_3\leqq X_k<-R_2\},\\
\nu_1^-&=\inf\{k>\mu_1^-\colon X_k\geqq -R_2\},\\
\mu_2^-&=\inf\{k>\nu_1^-\colon X_{k-1}\geqq -R_2,\ -R_3\leqq X_k<-R_2\},\\
\nu_2^-&=\inf\{k>\mu_2^-\colon X_k\geqq -R_2\},\\
&\ \vdots\,.
\end{aligned}$$

Note that if $\mu_m^+, \mu_m^-<k_n$, then we must have $m\leqq m_n$ (where $m_n\leqq(k_n-n)/2$). Hence if we let

$$D_n=\bigcup_{k=n}^{k_n-1}\{-R_4\leqq X_k\leqq R_2,\ X_{k+1}>R_3\}\cup\{-R_2\leqq X_k\leqq R_4,\ X_{k+1}<-R_3\},$$

$$E_n^+=\bigcup_{m=1}^{m_n}\{t_{\mu_m^+}<\sigma<t_{\nu_m^+},\ \sigma\leqq t_{k_n},\ \tau>t_{k_n}\},$$

$$E_n^-=\bigcup_{m=1}^{m_n}\{t_{\mu_m^-}<\sigma<t_{\nu_m^-},\ \sigma\leqq t_{k_n},\ \tau>t_{k_n}\},$$

then

$$P\{\sigma\leqq t_{k_n},\ \tau>t_{k_n}\}\leqq P\{D_n\}+P\{E_n^+\}+P\{E_n^-\}.$$

CLAIM 1. $\lim_{n\to\infty}P\{D_n\}=0$ *uniformly for* $|x|\leqq r$ *and all* x_0.

CLAIM 2. $\lim_{n\to\infty}P\{E_n^\pm\}=0$ *uniformly for* $|x|\leqq r$ *and all* x_0.

Assuming that Claims 1 and 2 hold, we have $P\{\sigma\leqq t_{k_n},\ \tau>t_{k_n}\}\to 0$ as $n\to\infty$. And the convergence is uniform for $|x|\leqq r$ and all x_0. This proves (3.8) and hence Lemma 4 when $d=1$.

Proof of Claim 1. Using the Chebyshev inequality and a standard estimate for the tail probability of a Gaussian random variable we have

$$
\begin{aligned}
P\{D_n\} &\leqq \sum_{k=n}^{k_n-1} P\{-R_4 \leqq X_k \leqq R_2,\, X_{k+1}-X_k > R_3-R_2\} \\
&\qquad \cup \{-R_2 \leqq X_k \leqq R_4,\, X_{k+1}-X_k < -(R_3-R_2)\} \\
&\leqq \sum_{k=n}^{k_n-1} P\{|X_k| \leqq R_4,\, |X_{k+1}-X_k| > R_3-R_2\} \\
&= \sum_{k=n}^{k_n-1} P\{|X_k| \leqq R_4,\, |-a_k(U'(X_k)+\xi_k)+b_k W_k| > R_3-R_2\} \\
&\leqq \sum_{k=n}^{k_n-1} \left(P\left\{|X_k| \leqq R_4,\, a_k|\xi_k| > \frac{R_3-R_2}{3}\right\} + P\left\{b_k|W_k| > \frac{R_3-R_2}{3}\right\}\right), \quad n \text{ large} \\
&\leqq c_1 \sum_{k=n}^{k_n-1} \left(a_k^2 E\{|\xi_k|^2,\, |X_k| \leqq R_4\} + \exp\left(-\frac{c_2}{b_k^2}\right)\right) \\
&\leqq c_3 \sum_{k=n}^{k_n-1} \left(a_k^{2+\alpha} + \exp\left(-\frac{c_2}{b_k^2}\right)\right) \\
&\leqq c_5 \sum_{k=n}^{\infty} \left(\frac{1}{k^{2+\alpha}} + \exp(-c_4 k)\right) \to 0 \quad \text{as } n \to \infty,
\end{aligned}
$$

since $\alpha > -1$. This completes the proof of Claim 1. □

Proof of Claim 2. Since the proofs for E_n^+ and E_n^- are symmetric, we only consider E_n^+. For convenience we suppress the + sign throughout, i.e., $E_n \triangleq E_n^+$, $\mu_m \triangleq \mu_m^+$, $\nu_m \triangleq \nu_m^+$.

For $1 \leqq m \leqq m_n$ let

$$E_{n,m} = \{t_{\mu_m} < \sigma < t_{\nu_m},\, \sigma \leqq t_{k_n},\, \tau > t_{k_n}\}.$$

We have

$$
\begin{aligned}
P\{E_{n,m}\} &= P \bigcup_{k=n+2}^{k_n} \{t_{\mu_m} < t_k < t_{\nu_m},\, \sigma = t_k,\, \tau > t_{k_n}\} \\
&= P \bigcup_{k=n+2}^{k_n} \{X_k - Y_k > R_4 - R_1,\, t_{\mu_m} < t_k < t_{\nu_m},\, \sigma = t_k,\, \tau > t_{k_n}\} \\
&\leqq P \bigcup_{k=n+2}^{k_n} \{X_k - Y_k > R_4 - R_1,\, t_{\mu_m} < t_k \leqq t_{\nu_m} \wedge \sigma \wedge \tau\} \\
&= P\left\{ \max_{k:\, t_{\mu_m} < t_k \leqq t_{\nu_m} \wedge \sigma \wedge \tau \wedge t_{k_n}} [X_k - Y_k] > R_4 - R_1 \right\} \\
&= P\left\{ \max_{k:\, t_{\mu_m} < t_k \leqq t_{\nu_m} \wedge \sigma \wedge \tau \wedge t_{k_n}} \left[X_{\mu_m} - Y_{\mu_m} - \sum_{l=\mu_m}^{k-1} a_l(U'(X_l) - U'(Y_l)) \right.\right. \\
&\qquad\qquad \left.\left. - \sum_{l=\mu_m}^{k-1} a_l(\xi_l - \zeta_l) \right] > R_4 - R_1 \right\}.
\end{aligned}
$$

Note that the $b_k W_k$ terms have been eliminated at this point; it is here we see the utility of comparing $X(t)$ and $Y(t)$. Now suppose $t_{\mu_m} < t_k \leqq t_{\nu_m} \wedge \sigma \wedge \tau \wedge t_{k_n}$. Then $X_{\mu_m} \in (R_2, R_3]$, $Y_{\mu_m} \in (-R_1, R_1)$, which implies $X_{\mu_m} - Y_{\mu_m} \leqq R_3 + R_1 = (R_4 - R_1)/2$.

Also $X_l \in (R_2, R_4]$, $Y_l \in (-R_1, R_1)$ for all l such that $\mu_m \leqq l < k$, which implies $U'(X_l) - U'(Y_l) > 0$ for all l such that $\mu_m \leqq l < k$. Now let

$$\eta_k = (\xi_k - \zeta_k)\mathbf{1}_{\{|X_k| \leqq R_4\}}.$$

Note that by (A4′) and Corollary 1

$$E\{|\eta_k|^2 | \mathscr{F}_k\} \leqq c_1 a_k^{\alpha \wedge 1}, \quad |E\{\eta_k | \mathscr{F}_k\}| \leqq c_1 a_k^{\beta \wedge (1/2)} \quad \text{w.p.1.}$$

Hence

$$\begin{aligned}
P\{E_{n,m}\} &\leqq P\left\{ \max_{k: t_{\mu_m} < t_k \leqq t_{\nu_m} \wedge \sigma \wedge \tau \wedge t_{k_n}} \sum_{l=\mu_m}^{k-1} a_l \eta_l > \frac{R_4 - R_1}{2} \right\} \\
&\leqq P\left\{ \max_{\mu_m < k \leqq \nu_m \wedge k_n} \sum_{l=\mu_m}^{k-1} a_l \eta_l > \frac{R_4 - R_1}{2} \right\} \\
&= P\left\{ \max_{n+1 \leqq k \leqq k_n - 1} \sum_{l=n+1}^{k} a_l \eta_l \mathbf{1}_{\{\mu_m \leqq l < \nu_m\}} > \frac{R_4 - R_1}{2} \right\} \\
&\leqq P\left\{ \max_{n+1 \leqq k \leqq k_n - 1} \sum_{l=n+1}^{k} a_l (\eta_l - E\{\eta_l | \mathscr{F}_l\}) \mathbf{1}_{\{\mu_m \leqq l < \nu_m\}} \right. \\
&\qquad \left. + \max_{n+1 \leqq k \leqq k_n - 1} \sum_{l=n+1}^{k} a_l E\{\eta_l | \mathscr{F}_l\} \mathbf{1}_{\{\mu_m \leqq l < \nu_m\}} > \frac{R_4 - R_1}{2} \right\}.
\end{aligned} \tag{3.9}$$

But

$$\begin{aligned}
\max_{n+1 \leqq k \leqq k_n - 1} &\left| \sum_{l=n+1}^{k} a_l E\{\eta_l | \mathscr{F}_l\} \mathbf{1}_{\{\mu_m \leqq l < \nu_m\}} \right| \\
&\leqq \sum_{l=n+1}^{k_n - 1} a_l |E\{\eta_l | \mathscr{F}_l\}| \\
&\leqq c_1 \sum_{l=n+1}^{k_n - 1} a_l^{(3/2) \wedge (1+\beta)} \\
&\leqq c_2 \sum_{l=n+1}^{\infty} \frac{1}{k^{(3/2) \wedge (1+\beta)}} \to 0 \quad \text{as } n \to \infty
\end{aligned} \tag{3.10}$$

since $\beta > 0$. Combining (3.9) and (3.10) gives for n large enough

$$P\{E_{n,m}\} \leqq P\left\{ \max_{n+1 \leqq k \leqq k_n - 1} \sum_{l=n+1}^{k} a_l (\eta_l - E\{\eta_l | \mathscr{F}_l\}) \mathbf{1}_{\{\mu_m \leqq l < \nu_m\}} > \frac{R_4 - R_1}{4} \right\}. \tag{3.11}$$

Let $\tilde{\eta}_k = \eta_k - E\{\eta_k | \mathscr{F}_k\}$ and

$$S_{m,k} = \sum_{l=n+1}^{k} a_l \tilde{\eta}_l \mathbf{1}_{\{\mu_m \leqq l < \nu_m\}}, \qquad k \geqq n+1.$$

Since $\tilde{\eta}_l$ is $\mathscr{F}_{l+1}$-measurable and $\{\mu_m \leqq l < \nu_m\} \in \mathscr{F}_l$, $\{S_{m,k}, \mathscr{F}_{k+1}\}_{k \geqq n+1}$ is a martingale. Hence applying Doob's inequality to (3.11) gives for n large enough

$$\begin{aligned}
P\{E_{n,m}\} &\leqq P\left\{ \max_{n+1 \leqq k \leqq k_n - 1} S_{m,k} > \frac{R_4 - R_1}{4} \right\} \\
&\leqq c_3 E\{S_{m,k_n - 1}^2\} \\
&= c_3 \sum_{k=n+1}^{k_n - 1} a_k^2 E\{|\tilde{\eta}_k|^2 \mathbf{1}_{\{\mu_m \leqq k < \nu_m\}}\}.
\end{aligned}$$

Finally,

$$
\begin{aligned}
P\{E_n\} &\leqq \sum_{m=1}^{m_n} P\{E_{n,m}\} \\
&\leqq c_3 \sum_{k=n+1}^{k_n-1} a_k^2 E\left\{|\tilde{\eta}_k|^2 \sum_{m=1}^{m_n} \mathbf{1}_{\{\mu_m \leqq k < \nu_m\}}\right\} \\
&\leqq c_3 \sum_{k=n+1}^{k_n-1} a_k^2 E\{|\tilde{\eta}_k|^2\} \\
&\leqq c_3 \sum_{k=n+1}^{k_n-1} a_k^2 E\{|\eta_k|^2\} \\
&\leqq c_4 \sum_{k=n+1}^{k_n-1} a_k^{3\wedge(2+\alpha)} \\
&\leqq c_5 \sum_{k=n+1}^{\infty} \frac{1}{k^{3\wedge(2+\alpha)}} \to 0 \quad \text{as } n \to \infty
\end{aligned}
$$

since $\alpha > -1$. This completes the proof of Claim 2.

Proof for $d > 1$. We now show how the above proof for $d = 1$ can be extended to $d > 1$.

Let u^i denote the ith component of a vector u. Suppose for the moment that there exists $R_2 > R_1$ such that for $R_3 = R_2 + 1$ and $R_4 = 2R_3 + 3R_1$, we have

$$
\sup_{\substack{x^i \leqq -R_2 \\ |x^j| \leqq R_4 \forall j \neq i}} \frac{\partial U}{\partial x^i}(x) < \inf_{|x^j| \leqq R_1 \forall j} \frac{\partial U}{\partial x^i}(x), \tag{3.12}
$$

$$
\inf_{\substack{x^i \geqq R_2 \\ |x^j| \leqq R_4 \forall j \neq i}} \frac{\partial U}{\partial x^i}(x) > \sup_{|x^j| \leqq R_1 \forall j} \frac{\partial U}{\partial x^i}(x). \tag{3.13}
$$

For $s > 0$, $R_0 > 0$, and $i = 1, \cdots, d$ let

$$
\sigma_i(s, R_0) = \inf\{t \geqq s \colon |X^i(t)| > R_0\}.
$$

Then we can show that as $n \to \infty$

$$
\begin{aligned}
&P_{0,x_0;t_n,x}\{\sigma(t_n, \sqrt{d}\, R_4) \leqq t_{k_n}, \tau(t_n, R_1) > t_{k_n}\} \\
&\qquad \leqq \sum_{i=1}^{d} P_{0,x_0;t_n,x}\{\sigma_i(t_n, R_4) \leqq t_{k_n}, \sigma_i(t_n, R_4) \leqq \sigma_j(t_n, R_4)\ \forall j \neq i, \tau(t_n, R_1) > t_{k_n}\} \to 0
\end{aligned}
$$

similarly to the proof given above that

$$
P_{0,x_0;t_n,x}\{\sigma(t_n, R_4) \leqq t_{k_n}, \tau(t_n, R_1) > t_{k_n}\} \to 0
$$

in the scalar case $d = 1$. So (3.8) and hence Lemma 4 holds for $R = \sqrt{d}\, R_4$.

It remains to establish (3.12) and (3.13). We only consider (3.13). Let $D(R_2) = \{x \colon x^i \geqq R_2, |x^j| \leqq R_4\ \forall j \neq i\}$. Since R_1 is fixed here, there will exist R_2 such that (3.13) holds if we can show

$$
\lim_{R_2 \to \infty} \inf_{x \in D(R_2)} \frac{\partial U}{\partial x^i}(x) = \infty.
$$

We proceed by breaking $\nabla U(x)$ into radial and tangential components and comparing the projection of these components on e^i, the ith standard basis element in $\mathbb{R}^d$. So let

$$\begin{aligned}
\hat{x} &= \frac{x}{|x|}, && |x| > 0, \\
\hat{\theta} &= \frac{\nabla U(x) - \langle \nabla U(x), \hat{x} \rangle \hat{x}}{|\nabla U(x) - \langle \nabla U(x), \hat{x} \rangle \hat{x}|}, && |\nabla U(x) - \langle \nabla U(x), \hat{x} \rangle \hat{x}| > 0 \\
&= 0, && |\nabla U(x) - \langle \nabla U(x), \hat{x} \rangle \hat{x}| = 0
\end{aligned}$$

and

$$g(x) = \frac{\langle \nabla U(x), \hat{\theta} \rangle \langle \hat{\theta}, e^i \rangle}{\langle \nabla U(x), \hat{x} \rangle \langle \hat{x}, e^i \rangle}, \quad x^i \text{ large.}$$

Of course $\langle \hat{x}, \hat{\theta} \rangle = 0$. Then

$$\begin{aligned}
\overline{\lim_{R_2 \to \infty}} \sup_{x \in D(R_2)} g^2(x) &\leqq \overline{\lim_{R_2 \to \infty}} \sup_{x \in D(R_2)} \frac{|\nabla U(x)|^2 - \langle \nabla U(x), \hat{x} \rangle^2}{\langle \nabla U(x), \hat{x} \rangle^2} \cdot \frac{1 - \langle \hat{x}, e^i \rangle^2}{\langle \hat{x}, e^i \rangle^2} \\
&= \overline{\lim_{R_2 \to \infty}} \sup_{x \in D(R_2)} \left[\left\langle \frac{\nabla U(x)}{|\nabla U(x)|}, \frac{x}{|x|} \right\rangle^{-2} - 1 \right] \cdot \frac{|x|^2 - (x^i)^2}{(x^i)^2} \\
&< (L(d)^{-2} - 1) 4(d-1) = 1,
\end{aligned}$$

where the first inequality follows from Bessel's inequality (applied to $\nabla U(x)$ and to e^i), and the last inequality follows from (A3) and the fact that if $x \in D(R_2)$ then $|x|^2 - (x^i)^2 \leqq (d-1) R_4^2$ and $(x^i)^2 \geqq R_2^2$ (and also $R_4 \sim 2R_2$ as $R_2 \to \infty$). Hence

$$\begin{aligned}
\underline{\lim_{R_2 \to \infty}} \inf_{x \in D(R_2)} \frac{\partial U}{\partial x^i}(x) &= \underline{\lim_{R_2 \to \infty}} \inf_{x \in D(R_2)} [\langle \nabla U(x), \hat{x} \rangle \langle \hat{x}, e^i \rangle + \langle \nabla U(x), \hat{\theta} \rangle \langle \hat{\theta}, e^i \rangle] \\
&= \underline{\lim_{R_2 \to \infty}} \inf_{x \in D(R_2)} \langle \nabla U(x), \hat{x} \rangle \langle \hat{x}, e^i \rangle (1 + g(x)) \\
&= \underline{\lim_{R_2 \to \infty}} \inf_{x \in D(R_2)} |\nabla U(x)| \left\langle \frac{\nabla U(x)}{|\nabla U(x)|}, \frac{x}{|x|} \right\rangle \frac{x^i}{|x|} (1 + g(x)) = \infty.
\end{aligned}$$

Hence (3.13) and similarly (3.12) follows. This completes the proof of Lemma 4.

3.2. Proof of Lemma 5. The idea behind this proof is that if $X(s) = Y(s)$ and $X(t)$ and $Y(t)$ remain in a fixed bounded set on large time intervals $t \in [s, \beta(s)]$ (and they do by Lemmas 3 and 4), then we can develop a recursion for estimating $E\{|X(\beta(s)) - Y(\beta(s))|^2\}$, and from the recursion we can show that $E\{|X(\beta(s)) - Y(\beta(s))|^2\} \to 0$ as $s \to \infty$. This is true even though the interval length $\beta(s) - s \to \infty$ as $s \to \infty$.

For each n let k_n be the integer that satisfies $\beta(t_n) \in [t_{k_n}, t_{k_{n+1}})$. We show that

$$(3.14) \qquad \lim_{n \to \infty} E_{0, x_0; t_n, x} \{|X(t_{k_n}) - Y(t_{k_n})|^2, \sigma(t_n, R) \wedge \tau(t_n, R) > t_{k_n}\} = 0.$$

The lemma then follows by some minor details, which are omitted.

In this proof we can and will assume that $\nabla U(\cdot)$ is bounded and Lipschitz function on $\mathbb{R}^d$, and ξ_k satisfies (A4′) with $K = \mathbb{R}^d$ (instead of K a compact subset of $\mathbb{R}^d$), i.e.,

$$(3.15) \qquad E\{|\xi_k|^2 | \mathscr{F}_k\} \leqq L a_k^\alpha, \quad |E\{\xi_k | \mathscr{F}_k\}| \leqq L a_k^\beta \quad \text{w.p.1}$$

(if $\sigma(t_n, R) \wedge \tau(t_n, R) > t_{k_n}$ then $|X(t)|$, $|Y(t)| \leqq R$ for $t_n \leqq t \leqq t_{k_n}$, and so $U(x)$ can be modified for $|x| > R$ and we can set $\xi_k = 0$ for $|X_k| > R$ without loss of generality).

Fix n for the moment and let $\sigma = \sigma(t_n, R)$, $\tau = \tau(t_n, R)$. Let

$$\zeta_k = \zeta(t_k, t_{k+1} \wedge \tau(t_k, R))$$

and

$$Y_{k+1} = Y_k - a_k(\nabla U(Y_k) + \zeta_k) + b_k W_k.$$

Note that if $Y(t_n) = Y_n$ and $\tau > t_{k_n}$, then $Y(t_{k_n}) = Y_{k_n}$. Henceforth assume all quantities are conditioned on $X(0) = X_0 = x_0$, $X(t_n) = X_n = Y(t_n) = Y_n = x$, $|x| \leqq r$. Then

$$\begin{aligned} E\{|X(t_{k_n}) - Y(t_{k_n})|^2, \sigma \wedge \tau > t_{k_n}\} &= E\{|X_{k_n} - Y_{k_n}|^2, \sigma \wedge \tau > t_{k_n}\} \\ &\leqq E\{|X_{k_n} - Y_{k_n}|^2\}. \end{aligned} \tag{3.16}$$

We proceed to show that the right side of (3.16) tends to zero as $n \to \infty$. Let

$$\Delta_k = X_k - Y_k, \qquad \eta_k = \xi_k - \zeta_k.$$

Note that by (3.15) and Corollary 1

$$E\{|\eta_k|^2 | \mathcal{F}_k\} \leqq c_1 a_k^{\alpha \wedge 1}, \quad |E\{\eta_k | \mathcal{F}_k\}| \leqq c_1 a_k^{\beta \wedge (1/2)} \quad \text{w.p.1.}$$

Now using Holder's inequality and the fact that X_k, Y_k, and hence Δ_k are $\mathcal{F}_k$ measurable we have

$$\begin{aligned} E\{|\Delta_{k+1}|^2\} &= E\{|\Delta_k - a_k(\nabla U(X_k + \Delta_k) - \nabla U(X_k) + \eta_k)|^2\} \\ &= E\{|\Delta_k|^2\} - 2a_k E\{\langle \Delta_k, \nabla U(X_k + \Delta_k) - \nabla U(X_k)\rangle\} \\ &\quad -2a_k E\{\langle \Delta_k, \eta_k\rangle\} + a_k^2 E\{|\nabla U(X_k + \Delta_k) - \nabla U(X_k)|^2\} \\ &\quad + 2a_k^2 E\{\langle \nabla U(X_k + \Delta_k) - \nabla U(X_k), \eta_k\rangle\} + a_k^2 E\{|\eta_k^2|\} \\ &\leqq E\{|\Delta_k|^2\} + 2d_1 a_k E\{|\Delta_k|^2\} \\ &\quad + 2a_k E\{|\Delta_k|^2\}^{1/2} E\{|E\{\eta_k|\mathcal{F}_k\}|^2\}^{1/2} + 2d_1^2 a_k^2 E\{|\Delta_k|^2\} \\ &\quad + 2d_1 a_k^2 E\{|\Delta_k|^2\}^{1/2} E\{|E\{\eta_k|\mathcal{F}_k\}|^2\}^{1/2} + a_k^2 E\{E\{|\eta_k|^2|\mathcal{F}_k\}\} \\ &\leqq (1 + c_2 a_k) E\{|\Delta_k|^2\} + c_2 a_k^\delta, \end{aligned}$$

where d_1 is a Lipschitz constant for $\nabla U(\cdot)$ and $\delta = \min[\frac{3}{2}, 2 + \alpha, 1 + \beta]$. Using the assumptions that $\alpha > -1$ and $\beta > 0$ we have $\delta > 1$. Now for each n

$$\begin{gathered} E\{|\Delta_{k+1}|^2\} \leqq (1 + c_2 a_k) E\{|\Delta_k|^2\} + c_2 a_k^\delta, \qquad k \geqq n, \\ E\{|\Delta_n|^2\} = 0, \end{gathered}$$

and if we replace the inequality with equality, the resulting difference equation is unstable as $k \to \infty$ (recall that $a_k = A/k$, k large). Nonetheless, we make the following claim.

CLAIM 3. *There exists $\gamma > 1$ such that*

$$\lim_{n \to \infty} \sup_{k: t_n \leqq t_k \leqq \gamma t_n} E\{|\Delta_k|^2\} = 0.$$

Assume the claim holds. Since $t_{k_n} \leqq \beta(t_n) \leqq t_n + 2t_n^{2/3} < \gamma t_n$ for n large, it follows that

$$\lim_{n \to \infty} E\{|\Delta_{k_n}|^2\} = 0.$$

This proves (3.14) and hence Lemma 5. It remains to prove the claim.

Proof of Claim 3. For each n let $\{u_{n,k}\}_{k\geqq n}$ be a sequence of nonnegative numbers such that

$$u_{n,k+1}\leqq(1+a_k)u_{n,k}+a_k^\delta,\qquad k\geqq n,$$
$$u_{n,n}=0,$$

where $\delta>1$. Now

$$u_{n,k}\leqq\sum_{m=n}^{k-1}a_m^\delta\prod_{l=m+1}^{k-1}(1+a_l)\leqq\left(\sum_{m=n}^{k-1}a_m^\delta\right)\cdot\exp\left(\sum_{m=n}^{k-1}a_m\right),$$

since $1+x\leqq e^x$. Also $\sum_n^{k-1}a_m\leqq A(\log(k/n)+1/n)$ and $\sum_n^{k-1}a_m^\delta\leqq A(1/(\delta-1)n^{\delta-1}+1/n^\delta)$, and if $t_k\leqq\gamma t_n$ then $k\leqq c_1n^\gamma$. Choose γ such that $1<\gamma<1+(\delta-1)/A$. It follows that

$$\sup_{k:t_n\leqq t_k\leqq\gamma t_n}u_{n,k}\leqq c_2n^{(\gamma-1)A-(\delta-1)}\to0\quad\text{as }n\to\infty.$$

The claim follows by setting $u_{n,k}=E\{|\Delta_k|^2\}$.

Remark. The proof of Claim 3 does *not* work if $a_k=A/k^\eta$ for any $\eta<1$.

4. General tightness criterion. In this section we consider the tightness of an algorithm of the form

$$X_{k+1}=X_k-a_k(\psi_k(X_k)+\xi_k)+b_kW_k,\qquad k\geqq0 \tag{4.1}$$

where $\{a_k\}$, $\{b_k\}$, $\{\xi_k\}$, and $\{W_k\}$ are defined as in § 2, and $\{\psi_k(x): x\in\mathbb{R}^d\}$ is an $\mathbb{R}^d$-valued random vector field for $k=0,1,\cdots$. We will deal with the following conditions in this section (α, β, γ_1, and γ_2 are constants whose values will be specified later).

(B1) For $k=0,1,\cdots$, let $\mathcal{F}_k=\sigma(X_0,W_0,\cdots,W_{k-1},\xi_0,\cdots,\xi_{k-1})$. There exists $L_1>0$ such that

$$E\{|\xi_k|^2|\mathcal{F}_k\}\leqq L_1a_k^\alpha,\quad|E\{\xi_k|\mathcal{F}_k\}|\leqq L_1a_k^\beta\quad\text{w.p.1}$$

W_k is independent of $\mathcal{F}_k$.

(B2) Let K be a compact subset of $\mathbb{R}^d$. There exists $L_2>0$ such that

$$E\{|\psi_k(x)|^2|\mathcal{F}_k\}\leqq L_2\quad\forall x\in K,\quad\text{w.p.1.}$$

(B3) There exists $L_3, R>0$ such that

$$E\{|\psi_k(x)||\mathcal{F}_k\}^2\geqq L_3\frac{|x|^2}{a_k^{\gamma_1}}\quad\forall|x|>R,\quad\text{w.p.1.}$$

(B4) There exists $L_4, R>0$ such that

$$E\{|\psi_k(x)|^2|\mathcal{F}_k\}\leqq L_4\frac{|x|^2}{a_k^{\gamma_2}}\quad\forall|x|>R,\quad\text{w.p.1.}$$

(B5) There exists $L_5, R>0$ such that

$$E\{\langle\psi_k(x),x\rangle|\mathcal{F}_k\}\geqq L_5E\{|\psi_k(x)||x||\mathcal{F}_k\}\quad\forall|x|>R,\quad\text{w.p.1.}$$

THEOREM 3. *Assume that* (B1)-(B5) *hold with* $\alpha>-1,\beta>0$, *and* $0\leqq\gamma_1\leqq\gamma_2<1$. *Let* $\{X_k\}$ *be given by* (4.1) *and* K *be a compact subset of* $\mathbb{R}^d$. *Then* $\{X_k^{x_0}: k\geqq0, x_0\in K\}$ *is a tight family of random variables.*

The proof of Theorem 3 will require the following lemmas.

LEMMA 6. *Assume the conditions of Theorem* 3. *Then there exist an integer* k_0 *and an* $M_1>0$ *such that*

$$E_{0,x_0}\{|X_{k+1}|^2\}-E_{0,x_0}\{|X_k|^2\}\leqq0\quad\textit{if } E_{0,x_0}\{|X_k|^2\}\geqq M_1,$$

for $k\geqq k_0$ *and all* x_0.

Proof. Assume all quantities are conditioned on $X_0 = x_0$. Now it follows from (B2)–(B5) and the fact that X_k is $\mathcal{F}_k$-measurable that

$$\begin{aligned}
&E\{|\psi_k(X_k)|^2, |X_k| \leqq R\} \leqq L_2,\\
&E\{|\psi_k(X_k)|^2, |X_k| > R\} \geqq L_3 a_k^{-\gamma_1} E\{|X_k|^2, |X_k| > R\},\\
&E\{|\psi_k(X_k)|^2, |X_k| > R\} \leqq L_4 a_k^{-\gamma_2} E\{|X_k|^2, |X_k| > R\},\\
&E\{\langle \psi_k(X_k), X_k\rangle, |X_k| > R\} \geqq L_5 L_3^{1/2} a_k^{-\gamma_1/2} E\{|X_k|^2, |X_k| > R\}.
\end{aligned}$$

Let $D \in \mathcal{F}_k$. Then using Holder's inequality and the fact that X_k is $\mathcal{F}_k$-measurable and W_k is independent of $\mathcal{F}_k$ we have

$$\begin{aligned}
&E\{|X_{k+1}|^2, D\} - E\{|X_k|^2, D\}\\
&\quad = E\{|X_k - a_k(\psi_k(X_k) + \xi_k) + b_k W_k|^2, D\} - E\{|X_k|^2, D\}\\
&\quad = -2a_k E\{\langle X_k, \psi_k(X_k)\rangle, D\} - 2a_k E\{\langle X_k, \xi_k\rangle, D\}\\
&\qquad + 2b_k E\{\langle X_k, W_k\rangle, D\} + a_k^2 E\{|\psi_k(X_k)|^2, D\}\\
&\qquad + 2a_k^2 E\{\langle \psi_k(X_k), \xi_k\rangle, D\} - 2a_k b_k E\{\langle \psi_k(X_k), W_k\rangle, D\}\\
&\qquad + a_k^2 E\{|\xi_k|^2, D\} - 2a_k b_k E\{\langle \xi_k, W_k\rangle, D\} + b_k^2 E\{|W_k|^2, D\}
\end{aligned}$$

$$\begin{aligned}
&\quad \leqq -2a_k E\{\langle X_k, \psi_k(X_k)\rangle, D\}\\
&\qquad + 2a_k E\{|X_k|^2, D\}^{1/2} E\{|E\{\xi_k | \mathcal{F}_k\}|^2\}^{1/2}\\
&\qquad + 2b_k E\{\langle X_k, E\{W_k\}\rangle, D\} + a_k^2 E\{|\psi_k(X_k)|^2, D\}\\
&\qquad + 2a_k^2 E\{|\psi_k(X_k)|^2, D\}^{1/2} E\{E\{|\xi_k|^2 | \mathcal{F}_k\}\}^{1/2}\\
&\qquad + 2a_k b_k E\{|\psi_k(X_k)|^2, D\}^{1/2} E\{|W_k|^2\}^{1/2}\\
&\qquad + a_k^2 E\{E\{|\xi_k|^2 | \mathcal{F}_k\}\}\\
&\qquad + 2a_k b_k E\{E\{|\xi_k|^2 | \mathcal{F}_k\}\}^{1/2} E\{|W_k|^2\}^{1/2} + b_k^2 E\{|W_k|^2\}.
\end{aligned} \tag{4.2}$$

Let $D = \{X_k > R\}$. Then using (4.2) we have

$$\begin{aligned}
&E\{|X_{k+1}|^2, |X_k| > R\} - E\{|X_k|^2, |X_k| > R\}\\
&\quad \leqq -c_1 a_k^{1-\gamma_1/2} E\{|X_k|^2, |X_k| > R\}\\
&\qquad + c_2((a_k^{\delta_1} + a_k^{1-\gamma_2/2} b_k) E\{|X_k|^2, |X_k| > R\} + a_k^{\delta_2} + a_k^{\delta_3} b_k + b_k^2),
\end{aligned}$$

where $\delta_1 = \min[1+\beta, 2-\gamma_2, 2+(\alpha-\gamma_2)/2]$, $\delta_2 = \min[1+\beta, 2+(\alpha-\gamma_2)/2, 2+\alpha]$, and $\delta_3 = \min[1-\gamma_2/2, 1+\alpha/2]$. Using the assumptions that $\alpha > -1$, $\beta > 0$, and $0 \leqq \gamma_1 \leqq \gamma_2 < 1$, we have $\delta_1 > 1$, $\delta_2 > 1$, and $\delta_3 > \frac{1}{2}$, and since $b_k = o(a_k^{1/2})$ we get

$$\begin{aligned}
&E\{|X_{k+1}|^2, |X_k| > R\} - E\{|X_k|^2, |X_k| > R\}\\
&\quad \leqq (-c_3 a_k^{1-\gamma_1/2} + o(a_k^{1-\gamma_1/2})) E\{|X_k|^2, |X_k| > R\} + o(a_k^{1-\gamma_1/2})\\
&\quad \leqq -c_4 a_k^{1-\gamma_1/2} (E\{|X_k|^2\} - R - 1)
\end{aligned} \tag{4.3}$$

for all $k \geqq k_0$, if we choose k_0 large enough.

Let $D = \{X_k \leqq R\}$. Then using (4.2) we have

$$E\{|X_{k+1}|^2, |X_k| \leqq R\} - E\{|X_k|^2, |X_k| \leqq R\} \leqq c_5(a_k^{\delta_4} + a_k^{\delta_5} b_k + b_k^2),$$

where $\delta_4 = \min[1, 1+\beta, 2+\alpha/2, 2+\alpha]$ and $\delta_5 = \min(1, 1+\alpha/2]$. Using the assumptions that $\alpha > -1$ and $\beta > 0$ we have $\delta_4 = 1$ and $\delta_5 > \frac{1}{2}$, and since $b_k = o(a_k^{1/2})$ we get

$$(4.4) \qquad E\{|X_{k+1}|^2, |X_k| \leqq R\} - E\{|X_k|^2, |X_k| \leqq R\} \leqq c_6 a_k \leqq c_6 a_k^{1-\gamma_1/2}$$

for all $k \geqq 0$.

Finally, let $M_1 = c_6/c_4 + R + 1$. Then combining (4.3) and (4.4) gives the lemma. □

LEMMA 7. *Assume the conditions of Theorem* 3. *Then there exists an* $M_2 > 0$ *such that*

$$E_{0,x_0}\{|X_{k+1}|^2\} - E_{0,x_0}\{|X_k|^2\} \leqq M_2(E_{0,x_0}\{|X_k|^2\} + 1)$$

for $k \geqq 0$ *and all* x_0.

Proof. Similarly to the proof of Lemma 6 we can show that conditioned on $X_0 = x_0$

$$E\{|X_{k+1}|^2, |X_k| > R\} - E\{|X_k|^2, |X_k| > R\} \leqq c_1 a_k^{1/2}(E\{|X_k|^2\} + 1)$$

and

$$E\{|X_{k+1}|^2, |X_k| \leqq R\} - E\{|X_k|^2, |X_k| \leqq R\} \leqq c_1 a_k.$$

Combining these equations gives the lemma. □

Proof of Theorem 3. Let M_1, M_2, and k_0 be as in Lemmas 6 and 7. By Lemma 7 there exists $c_1 \geqq M_1$ such that

$$E_{0,x_0}\{|X_k|^2\} \leqq c_1, \quad \forall k \leqq k_0, \quad x_0 \in K,$$

and by Lemmas 6 and 7 we also have

$$E_{0,x_0}\{|X_{k+1}|^2\} - E_{0,x_0}\{|X_k|^2\} \leqq 0 \quad \text{if } E_{0,x_0}\{|X_k|^2\} \geqq M_1$$

and

$$E_{0,x_0}\{|X_{k+1}|^2\} - E_{0,x_0}\{|X_k|^2\} \leqq M_2(E_{0,x_0}\{|X_k|^2\} + 1)$$

for $k \geqq k_0$ and all x_0. Let $c_2 = c_1 + M_2(M_1 + 1)$. Then by induction we get

$$E_{0,x_0}\{|X_k|^2\} \leqq c_2 \quad \forall k \geqq 0, \quad x_0 \in K,$$

and the tightness of $\{X_k^{x_0}: k \geqq 0, x_0 \in K\}$ follows from this. □

5. Tightness and convergence for two example algorithms. In this section we apply Theorems 2 and 3 to establish the tightness and ultimately the convergence of two example algorithms. Define $U(\cdot)$, $\{a_k\}$, $\{b_k\}$, $\{\xi_k\}$, and $\{W_k\}$ as in § 2. We will need to consider one or both of the following conditions:

$$(\text{A5}) \qquad \underline{\lim}_{|x| \to \infty} |\nabla U(x)|/|x| > 0.$$

$$(\text{A6}) \qquad \overline{\lim}_{|x| \to \infty} |\nabla U(x)|/|x| < \infty.$$

Example 1. Here we consider the following algorithm:

$$(5.1) \qquad X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k, \qquad k \geqq 0.$$

THEOREM 4. *Assume* (A1)-(A3), (B1), (A5), and (A6) *hold with* $\alpha > -1, \beta > 0$. *Let* $\{X_k\}$ *be given by* (5.1). *Then for* $B/A > C_0$ *and any bounded continuous function* $f(\cdot)$ *on* $\mathbb{R}^d$

$$\lim_{k \to \infty} E_{0,x_0}\{f(X_k)\} = \pi(f)$$

uniformly for x_0 *in a compact set.*

Proof. The assumptions of Theorem 2 and Theorem 3 (with $\psi_k(x) = \nabla U(x)$ and $\gamma_1 = \gamma_2 = 0$) are satisfied. □

Observe that the proof of tightness of $\{X_k^{x_0}\}$ using Theorem 3 requires that (A5) and (A6) hold, i.e., there exists M_1 and M_2 such that

$$M_1|x| \leqq |\nabla U(x)| \leqq M_2|x|, \quad |x| \text{ large.}$$

Intuitively, the upper bound on $|\nabla U(x)|$ is needed to prevent potentially unbounded oscillations of $\{X_k\}$ around the origin. It is possible to modify (5.1) in such a way that only the lower bound on $|\nabla U(x)|$ (i.e., (A5)) but not the upper bound on $|\nabla U(x)|$ (i.e., (A6)) is needed. Since we still want convergence to a global minimum of $U(\cdot)$, which is not known to lie in a specified bounded domain, standard multiplier and projection methods [1] are precluded. The next example gives a modification of (5.1), which has the desired properties.

Example 2. Here we consider the following algorithm:

$$(5.2) \qquad \begin{aligned} X_{k+1} &= X_k - a_k(\nabla U(X_k) + \xi_k) + b_k W_k \quad && \text{if} \quad |\nabla U(X_k) + \xi_k| \leqq \frac{|X_k| \vee 1}{a_k^\gamma} \\ &= X_k - a_k^{1-\gamma} X_k + b_k W_k && \text{if} \quad |\nabla U(X_k) + \xi_k| > \frac{|X_k| \vee 1}{a_k^\gamma}, \end{aligned}$$

where $\gamma > 0$. Intuitively, note that if K is a fixed compact set, $X_k \in K$, ξ_k is not too large, and k is very large, then X_k is updated to X_{k+1} as in (5.1). Also note that in (5.2) (like (5.1)), $\nabla U(X_k)$ and ξ_k only appear as the sum $\nabla U(X_k) + \xi_k$. This means that we can use noisy or imprecise measurements of $\nabla U(\cdot)$ in (5.2) in exactly the same way as in (5.1).

THEOREM 5. *Assume* (A1)-(A3), (B1), *and* (A5) (*but not necessarily* (A6)) *hold with* $\alpha > 0$. *Let* $\{X_k\}$ *be given by* (5.2) *with* $0 < \gamma < \frac{1}{2}$. *Then for* $B/A > C_0$ *and any bounded continuous function* $f(\cdot)$ *on* $\mathbb{R}^d$

$$(5.3) \qquad \lim_{k \to \infty} E_{0,x_0}\{f(X_k)\} = \pi(f)$$

uniformly for x_0 *in a compact set.*

Proof. Let

$$(5.4) \qquad X_{k+1} = X_k - a_k(\nabla U(X_k) + \xi_k') + b_k W_k$$

(this defines ξ_k') and $\mathscr{F}_k' = \sigma(X_0, \xi_0', \cdots, \xi_{k-1}', W_0, \cdots, W_{k-1})$. We show that $(\xi_k', W_k, \mathscr{F}_k')$ satisfies (A4). Hence by Theorem 2 if $\{X_k^{x_0}: k \geqq 0, x_0 \in K\}$ is tight for K compact then (5.3) holds.

Let

$$\begin{aligned} \psi_k(x) &= \nabla U(x) \quad && \text{if} \quad |\nabla U(x) + \xi_k| \leqq \frac{|x| \vee 1}{a_k^\gamma} \\ &= \frac{x}{a_k^\gamma} && \text{if} \quad |\nabla U(x) + \xi_k| > \frac{|x| \vee 1}{a_k^\gamma}. \end{aligned}$$

Let

$$(5.5) \qquad X_{k+1} = X_k - a_k(\psi_k(X_k) + \xi_k'') + b_k W_k$$

(this defines ξ_k'') and $\mathscr{F}_k'' = \sigma(X_0, \xi_0'', \cdots, \xi_{k-1}'', W_0, \cdots, W_{k-1})$. We show that $(\xi_k'', W_k, \mathscr{F}_k'')$ satisfies (B1) and $(\psi_k(x), \mathscr{F}_k'')$ satisfies (B2)-(B5) with $\gamma_1 = 0$, $\gamma_2 = 2\gamma$. Hence by Theorem 3 $\{X_k^{x_0}: k \geqq 0, x_0 \in K\}$ is tight for K compact and (5.3) does hold. These assertions are proved in Claims 4 and 5 below.

Remark. The proof shows the importance of separating the tightness and convergence issues. Even though we can write algorithm (5.2) in the form of algorithm (5.4), we cannot apply Theorem 3 to (5.4) to prove tightness because $U(\cdot)$ may not satisfy (A6), and ξ'_k may not satisfy (B1) even though ξ_k satisfies (B1).

CLAIM 4. *Let K be a compact subset of $\mathbb{R}^d$. Then there exists $M_1 > 0$ such that*

$$E\{|\xi'_k|^2|\mathcal{F}'_k\} \leqq M_1 a_k^\alpha \quad \forall X_k \in K, \quad w.p.1.$$

Also, W_k is independent of $\mathcal{F}'_k$.

Proof. Clearly,

$$\begin{aligned} \xi'_k &= \xi_k && \text{if} \quad |\nabla U(X_k) + \xi_k| \leqq \frac{|X_k| \vee 1}{a_k^\gamma} \\ &= \frac{X_k}{a_k^\gamma} - \nabla U(X_k) && \text{if} \quad |\nabla U(X_k) + \xi_k| > \frac{|X_k| \vee 1}{a_k^\gamma}. \end{aligned}$$

Hence for $X_k \in K$ and k large enough

$$\begin{aligned} E\{|\xi'_k|^2|\mathcal{F}_k\} &\leqq E\{|\xi_k|^2|\mathcal{F}_k\} + E\left\{\left|\frac{X_k}{a_k^\gamma} - \nabla U(X_k)\right|^2, |\nabla U(X_k) + \xi_k| > \frac{|X_k| \vee 1}{a_k^\gamma} \,\middle|\, \mathcal{F}_k\right\} \\ &\leqq L_1 a_k^\alpha + \frac{c_1}{a_k^{2\gamma}} \Pr\left\{|\nabla U(X_k) + \xi_k| > \frac{|X_k| \vee 1}{a_k^\gamma} \,\middle|\, \mathcal{F}_k\right\} \\ &\leqq L_1 a_k^\alpha + \frac{c_1}{a_k^{2\gamma}} \Pr\left\{|\xi_k| > \frac{c_2}{a_k^\gamma} \,\middle|\, \mathcal{F}_k\right\} \\ &\leqq L_1 a_k^\alpha + c_3 E\{|\xi_k|^2|\mathcal{F}_k\} \leqq M_1 a_k^\alpha \quad \text{w.p.1}, \end{aligned}$$

where we have used the assumption that $\gamma > 0$ and the Chebyshev inequality. It is easy to see that the inequality actually holds for all $k \geqq 0$. Since $\mathcal{F}'_k \subset \mathcal{F}_k$, the claim follows.

CLAIM 5. *Let K be a compact subset of $\mathbb{R}^d$. Then there exists M_1, M_2, M_3, M_4, and M_5, $R > 0$ such that*

(i) *$E\{|\xi''_k|^2|\mathcal{F}''_k\} \leqq M_1 a_k^\alpha$ w.p.1. Also W_k is independent of $\mathcal{F}''_k$,*

(ii) *$E\{|\psi_k(x)|^2|\mathcal{F}''_k\} \leqq M_2$ for all $x \in K$, w.p.1,*

(iii) *$E\{|\psi_k(x)||\mathcal{F}''_k\}^2 \geqq M_3|x|^2$ for all $|x| > R$, w.p.1,*

(iv) *$E\{|\psi_k(x)|^2|\mathcal{F}''_k\} \leqq M_4(|x|^2/a_k^{2\gamma})$ for all $|x| > R$, w.p.1,*

(v) *$E\{\langle\psi_k(x), x\rangle|\mathcal{F}''_k\} \geqq M_5 E\{|\psi_k(x)||x||\mathcal{F}''_k\}$ for all $|x| > R$, w.p.1.*

Proof. First observe that (iii) and (v) follow immediately from (A3) and (A5).

(i) Clearly,

$$\begin{aligned} \xi''_k &= \xi_k \quad \text{if} \quad |\nabla U(X_k) + \xi_k| \leqq \frac{|X_k| \vee 1}{a_k^\gamma} \\ &= 0 \quad \text{if} \quad |\nabla U(X_k) + \xi_k| > \frac{|X_k| \vee 1}{a_k^\gamma}. \end{aligned}$$

Hence

$$E\{|\xi''_k|^2|\mathcal{F}_k\} \leqq E\{|\xi_k|^2|\mathcal{F}_k\} \leqq M_1 a_k^\alpha \quad \text{w.p.1}.$$

Since $\mathcal{F}''_k \subset \mathcal{F}_k$, (i) must hold.

(ii) For $x \in K$ and k large enough

$$\begin{aligned} E\{|\psi_k(x)|^2|\mathscr{F}_k\} &\leqq |\nabla U(x)|^2 + \frac{|x|^2}{a_k^{2\gamma}} \Pr\left\{|\nabla U(x)+\xi_k| > \frac{|x| \vee 1}{a_k^{\gamma}} \,\middle|\, \mathscr{F}_k\right\} \\ &\leqq c_1 + \frac{c_1}{a_k^{2\gamma}} \Pr\left\{|\xi_k| > \frac{c_2}{a_k^{\gamma}} \,\middle|\, \mathscr{F}_k\right\} \\ &\leqq c_1 + c_3 E\{|\xi_k|^2|\mathscr{F}_k\} \leqq M_2 \quad \text{w.p.1}, \end{aligned}$$

where we have used the assumption that $\gamma > 0$ and the Chebyshev inequality. It is easy to see that the inequality actually holds for $x \in K$ and all $k \geqq 0$. Since $\mathscr{F}_k'' \subset \mathscr{F}_k$, (ii) must hold.

(iv) For $|x|$ large enough and $k \geqq 0$

$$\begin{aligned} E\{|\psi_k(x)|^2|\mathscr{F}_k\} &\leqq \frac{|x|^2}{a_k^{2\gamma}} + |\nabla U(x)|^2 P\left\{|\nabla U(x)+\xi_k| \leqq \frac{|x| \vee 1}{a_k^{\gamma}} \,\middle|\, \mathscr{F}_k\right\} \\ &\leqq \frac{|x|^2}{a_k^{2\gamma}} + E\left\{|\nabla U(x)|^2, |\nabla U(x)| \leqq \frac{|x|}{a_k^{\gamma}} + |\xi_k| \,\middle|\, \mathscr{F}_k\right\} \\ &\leqq 4\frac{|x|^2}{a_k^{2\gamma}} + 3E\{|\xi_k|^2|\mathscr{F}_k\} \leqq M_4 \frac{|x|^2}{a_k^{2\gamma}} \quad \text{w.p.1}. \end{aligned}$$

Since $\mathscr{F}_k'' \subset \mathscr{F}_k$, (iv) must hold. This completes the proof of the claim and hence the theorem. □

As a final note observe that the algorithm (5.1) does require (A6), and also (B1) with $\alpha > -1$, $\beta > 0$. On the other hand, the algorithm (5.2) does not require (A6), but does require (B1) with $\alpha > 0$ (and hence $\beta > 0$ by Holder's inequality). It may be possible to allow $\{\xi_k\}$ with unbounded variance in (5.2) but this would require some additional assumptions on $\{\xi_k\}$ and we do not pursue this.

REFERENCES

[1] H. J. KUSHNER AND D. CLARK, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Applied Mathematical Science Series 26, Springer-Verlag, Berlin, 1978.

[2] H. J. 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.

[3] L. LJUNG, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, 22 (1977), pp. 551-575.

[4] U. GRENENDER, *Tutorial in pattern theory*, Division of Applied Mathematics, Brown University, Providence, RI, 1984.

[5] S. GEMAN AND C. R. HWANG, *Diffusions for global optimization*, SIAM J. Control Optim., 24 (1986), pp. 1031-1043.

[6] B. GIDAS, *Global optimization via the Langevin equation*, in Proc. IEEE Conference on Decision and Control, Fort Lauderdale, FL, 1985, pp. 774-778.

[7] 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. 737-752.

[8] L. C. W. DIXON AND G. P. SZEGO, *Towards Global Optimization*, North-Holland, Amsterdam, 1978.

[9] C.-R. HWANG, *Laplace's method revisited: weak convergence of probability measures*, Ann. Probab., 8 (1980), pp. 1177-1182.

[10] M. I. FREIDLIN AND A. D. WENTZELL, *Random Perturbations of Dynamical Systems*, Springer-Verlag, Berlin, New York, 1984.

[11] J. L. DOOB, *Stochastic Processes*, John Wiley, New York, 1952.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 5, pp. 1019-1036, September 1991

ON SUBDIFFERENTIALS OF OPTIMAL VALUE FUNCTIONS*

LIONEL THIBAULT†

Abstract. A general mathematical programming problem in which the constraints are defined by multifunctions and depend on a parameter u, and the resulting value function $m(u)$ are considered. In the context of Banach spaces admitting equivalent Fréchet differentiable norms estimates for the generalized gradient ∂m of m are established. A special study is made of problems in which the multifunctions defining the constraints take convex values. For these problems, estimates for ∂m are given in terms of the generalized gradients of the support functions of these multifunctions.

Key words. optimal value function, subdifferential, generalized gradient, ε-Fréchet subgradient, singular subdifferential, support function, regular multifunction

AMS(MOS) subject classifications. 90C30, 90C31, 90C48

Introduction. The study of the behaviour of an optimal value function is known to be very important for the interpretation of marginal values for resources in an economic framework and for its direct significance in stability and sensitivity analysis. A useful tool that furnishes information about the behaviour of the optimal value function is its Clarke subdifferential. So, many authors have studied the Clarke subdifferential of the optimal value function of a problem depending on a parameter and some have proved that this subdifferential is related to the Lagrange multipliers (relative to necessary optimality conditions in terms of Clarke generalized gradients) of the concerned problem. In the case of problems with explicit constraints, this was probably begun with the paper of Gauvin [14]. Recently, a great deal of progress has been realized in this domain by Rockafellar. Indeed, in a series of papers [28]-[32], Rockafellar considered the following general finite-dimensional problem:

$$(O_u) \qquad \underset{x\in\mathbb{R}^n}{\text{minimize}}\ \{f(x,u): (x,u)\in D,\ g(x,u)\in C\},$$

where $D\subset\mathbb{R}^n\times\mathbb{R}^m$ and $C\subset\mathbb{R}^k$, f is an extended real-valued lower semicontinuous function, and g is a locally Lipschitz mapping (not necessarily differentiable). For this general problem, Rockafellar established estimates for the Clarke subdifferential of the value function in terms of Lagrange multipliers vectors that satisfy necessary conditions for (O_u) in terms of Clarke subdifferentials. He also proved that the already-known bounds on Dini derivatives of the value function follow from his subgradient estimates, without the restrictions on (O_u), which were often made in the past. All the results of Rockafellar in these papers are strongly related to the notion of proximal subgradients that he introduced in [28] and the exact formula giving the Clarke subdifferential of a function in terms of its limiting proximal subgradients in finite dimension. Let us note that Clarke [7] also furnished other proofs of estimates of subdifferentials of some general value functions in finite dimension by using the expression of the Clarke normal cones in terms of limiting perpendicular vectors in finite dimension and that with Clarke [7] and Clarke and Loewen [8], a great deal of progress has also been made in perturbed optimal control problems.

The present paper studies the Clarke subdifferential of the optimal value function of a general perturbed problem with a constraint defined by a multifunction:

$$m(u)=\min_x\{f(x,u): u\in M(x), x\in A\}.$$

* Received by the editors December 7, 1988; accepted for publication (in revised form) August 10, 1990.

† Département de Mathématiques, Faculté des Sciences, 64000 Pau, France.

In fact, we are interested in finding estimates for the Clarke subdifferential $\partial m(u)$ of m in terms of the coderivative of M (and the subdifferential of Δ_M, where $\Delta_M(x, y) = d(y, M(x))$) and the subdifferential of f, on the one hand, and mainly (whenever M takes convex values) in terms of the subdifferential of f and the subdifferential of the support function of M, on the other hand. To this end we follow the work begun by Rockafellar [31]. The nice result of Treiman [34], [35], giving an exact expression of the Clarke subdifferential in terms of limiting ε-Fréchet subgradients, and the general and no less deep calculus rules established by Kruger [20] and Kruger and Mordukhovich [21] for the generalized differentials introduced in [21], allow us to work in the context of Banach spaces admitting an equivalent norm that is Fréchet differentiable off zero (for example, any reflexive Banach space). By this method, at the same time we obtain estimates for the subdifferential $\delta m(u)$ in the sense of Kruger and Mordukhovich.

Sections 1 and 2 are devoted to recalling and to establishing some results that will be used throughout the paper. In § § 3 and 4 we give estimates for the Kruger-Mordukhovich subdifferentials and singular subdifferentials of m in terms of the subdifferential of f and the coderivative of M. By the result of Treiman estimates of the Clarke subdifferential of m can easily be derived. The particular and important case where M takes convex values is considered in the last section. Following Clarke [6], [7] we prove, for unperturbed problems with constraints defined by convex-valued multifunctions, first-order necessary optimality conditions (complementary to those of Dien [10], [11]) in terms of the subdifferentials of the support functions of these multifunctions. Then we give estimates for the subdifferential of m in terms of the subdifferentials of f and the support function of M.

Before closing this Introduction, let us indicate that two extensive lists of references on directional derivatives or subdifferentials of optimal value functions can be found in [7] and [29].

1. Distance function associated with a multifunction. Let M be a multifunction from a metric space E into a metric space F. We consider the function Δ_M defined on $E \times F$ by

$$\Delta_M(x, y) = d(y, M(x)),$$

where the right term is $-\infty$ whenever $M(x) = \varnothing$. As several estimates in this paper will be given in terms of subdifferentials of Δ_M, we show in this section the importance of this function in Clarke subdifferential theory.

Obviously,

$$d((x, y), GrM) \leqq \Delta_M(x, y) \quad \text{for all } (x, y) \in E \times F,$$

where $GrM = \{(x, y) \in E \times F : y \in M(x)\}$ and the distance on $E \times F$ is defined by

$$d((x, y), (x', y')) = (d(x, x')^2 + d(y, y')^2)^{1/2}.$$

In order to give a type of reverse inequality, let us recall the notion of pseudo-Lipschitz multifunction (see Aubin and Ekeland [1], [2]).

DEFINITION 1.1. Let $l \geqq 0$. M is said to be *l-pseudo-Lipschitz at* $(\bar{x}, \bar{y}) \in GrM$ if there are $X \in \mathcal{N}(\bar{x})$, $Y \in \mathcal{N}(\bar{y})$, such that

$$(1.1) \qquad Y \cap M(x) \subset \bigcup_{y \in M(x')} B(y, l d(x, x')) \quad \text{for all } x, x' \in X$$

(here $\mathcal{N}(\bar{x})$ denotes the filter of neighbourhoods of $\bar{x}$).

Remark. Rockafellar [32] has proved that M is l-pseudo-Lipschitz at $(\bar{x}, \bar{y})$ if and only if there are $X \in \mathcal{N}(\bar{x})$ and $Y \in \mathcal{N}(\bar{y})$ such that

$$(1.2) \qquad d(y, M(x)) \leqq d(y', M(x')) + d(y, y') + ld(x, x')$$

for all $(x, y), (x', y') \in (X \times Y)$.

The following result has already been obtained by Clarke [5] for Lipschitz multifunctions. Here we extend it to pseudo-Lipschitz multifunctions. Throughout the paper $B(x, r)$ will be the closed ball centered at x.

PROPOSITION 1.2. *If M is l-pseudo-Lipschitz at $(\bar{x}, \bar{y}) \in GrM$, then there exist $X \in \mathcal{N}(\bar{x})$, $Y \in N(\bar{y})$ such that*

$$\Delta_M(x, y) \leqq Kd((x, y), GrM) \quad \textit{for all } (x, y) \in X \times Y,$$

where $K = 2 \max(1, l)$.

Proof. Choose $\varepsilon > 0$ such that $B((\bar{x}, \bar{y}), 3\varepsilon) \subset X \times Y$, where X and Y are given by (1.2). Fix $(x, y) \in B((\bar{x}, \bar{y}), \varepsilon)$. For any $(x', y') \in E \times F$ with $d((x', y'), (\bar{x}, \bar{y})) > 3\varepsilon$ we have

$$\begin{aligned} d((x, y), (x', y')) &\geqq d((\bar{x}, \bar{y}), (x', y')) - d((\bar{x}, \bar{y}), (x, y)) \\ &\geqq 3\varepsilon - d((\bar{x}, \bar{y}), (x, y)) \\ &\geqq \varepsilon + d((x, y), (\bar{x}, \bar{y})) \geqq \varepsilon + d((x, y), GrM). \end{aligned}$$

So we have

$$d((x, y), GrM) = \inf \{(d((x, y), (x', y')) \colon (x', y') \in B((\bar{x}, \bar{y}), 3\varepsilon) \cap GrM\}.$$

But for any $(x', y') \in B((\bar{x}, \bar{y}), 3\varepsilon) \cap GrM$

$$\begin{aligned} d(y, M(x)) &\leqq d(y, M(x')) + ld(x, x') \\ &\leqq d(y, y') + ld(x, x') \\ &\leqq Kd((x, y), (x', y')), \end{aligned}$$

where $K = 2 \max(1, l)$ and hence

$$\Delta_M(x, y) \leqq Kd((x, y), GrM). \qquad \square$$

In the sequel E and F will be two real *normed vector spaces.* Let f be a function from E into $\mathbb{R} \cup \{-\infty, +\infty\}$ with $|f(\bar{x})| < \infty$. The generalized directional derivative $f^{\uparrow}(\bar{x}; \cdot)$ is defined (see Rockafellar [27]) by

$$\begin{aligned} f^{\uparrow}(\bar{x}; h) &= \limsup_{\substack{(x,\alpha)\downarrow_f \bar{x} \\ t \downarrow 0}} \inf_{h' \to h} t^{-1}[f(x + th') - \alpha] \\ &:= \sup_{H \in \mathcal{N}(h)} \left[\limsup_{\substack{(x,\alpha)\downarrow_f \bar{x} \\ t \downarrow 0}} \left(\inf_{h' \in H} t^{-1}(f(x + th') - \alpha) \right) \right] \end{aligned}$$

and the directional hyperderivative $f^0(\bar{x}; \cdot)$ by

$$f^0(\bar{x}; h) = \limsup_{\substack{(x,\alpha)\downarrow_f \bar{x} \\ t \downarrow 0}} t^{-1}[f(x + th) - \alpha],$$

where $(x, \alpha)\downarrow_f \bar{x}$ means

$$(x, \alpha) \in \operatorname{epi} f := \{(z, \beta) \in E \times \mathbb{R} \colon f(x) \leqq \beta\} \quad \text{and} \quad (x, \alpha) \to (\bar{x}, f(\bar{x})).$$

If f is Lipschitz around $\bar{x}$, then

$$f^{\uparrow}(\bar{x}; h) = f^0(\bar{x}; h) = \limsup_{\substack{x \to \bar{x} \\ t \downarrow 0}} t^{-1}[f(x+th) - f(x)].$$

Using a technique inspired by a method introduced by Hiriart-Urruty in [16] we have Proposition 1.3.

PROPOSITION 1.3. *Let M be any multifunction from E into F with $(\bar{x}, \bar{y}) \in GrM$. Then for all $(h, k) \in E \times F$*

$$\Delta_M^{\uparrow}(\bar{x}, \bar{y}; h, k) = \limsup_{\substack{(x,y) \xrightarrow[t \downarrow 0]{M} (\bar{x},\bar{y})}} \inf_{(h',k') \to (h,k)} t^{-1}[\Delta_M(x+th', y+tk')]$$

and

$$\Delta_M^0(\bar{x}, \bar{y}; h, k) = \limsup_{\substack{(x,y) \xrightarrow[t \downarrow 0]{M} (\bar{x},\bar{y})}} t^{-1}[\Delta_M(x+th, y+tk)],$$

where $(x, y) \xrightarrow{M} (\bar{x}, \bar{y})$ means $(x, y) \in GrM$ and $(x, y) \to (\bar{x}, \bar{y})$.

Proof. We only show the first equality (the second one is obtained with $H \times K = \{(h, k)\}$ in the proof below). It is enough to prove that the left-hand side of this equality is not greater than the other one (which we denote by β) since the reverse inequality is obviously true. Consider $\gamma > 0$ and $H \times K \in \mathcal{N}(h, k)$. There exists $X \times Y' \in N(\bar{x}, \bar{y})$, $\varepsilon' \in]0, 1[$ such that for all $t \in]0, \varepsilon'[$, $(x, y) \in (X \times Y') \cap GrM$ there is $(h', k') \in H \times K$ with (we suppose $\beta < +\infty$)

$$(1.3) \qquad t^{-1}\Delta_M(x+th', y+tk') < \beta + \gamma.$$

Choose $\varepsilon \in]0, \min(\varepsilon', \gamma)[$ and $Y \in \mathcal{N}(\bar{y})$ such that $Y + B(0, 2\varepsilon) \subset Y'$ and fix any $t \in]0, \varepsilon[$ and any $(x, y, \alpha) \in \operatorname{epi} \Delta_M \cap X \times Y \times]-\varepsilon, \varepsilon[$. As $\alpha \geqq \Delta_M(x, y) = d(y, M(x))$ we have $M(x) \neq \emptyset$ and hence we can choose $y' \in M(x)$

$$\|y - y'\| < t^2 + d(y, M(x)) < \varepsilon^2 + \alpha < 2\varepsilon,$$

and hence $y' \in y + B(0, 2\varepsilon) \subset Y'$, which ensures that $(x, y') \in (X \times Y') \cap GrM$. Therefore by (1.3) there exists $(h', k') \in H \times K$ such that

$$\begin{aligned} \beta + 2\gamma &> t^{-1}[\Delta_M(x+th', y'+tk')] + t \\ &\geqq t^{-1}[\Delta_M(x+th', y'+tk') + \|y - y'\| - \Delta_M(x, y)] \\ &\geqq t^{-1}[\Delta_M(x+th', y+tk') - \alpha], \end{aligned}$$

which proves that

$$\beta \geqq \Delta_M^{\uparrow}(\bar{x}, \bar{y}; h, k). \qquad \square$$

The function Δ_M has been sucessfully used by Clarke [5], [7] for Lipschitz multifunctions in optimal control theory. The following corollary makes clear that Δ_M can also be a powerful tool in optimization theory even for M not necessarily Lipschitz.

Before stating this corollary let us recall that the Clarke tangent cone $T(A; \bar{x})$ to a subset $A \subset E$ at a point $\bar{x} \in A$ is the set of all vectors $h \in E$ such that for all $x_n \xrightarrow{A} \bar{x}$, $t_n \downarrow 0$ there exists $h_n \to h$ in E with $x_n + t_n h_n \in A$ for all n. We know (see [7] and [27]) that

$$d_A^0(\bar{x}; h) = 0 \Leftrightarrow h \in T(A; \bar{x}) \Leftrightarrow \psi_A^{\uparrow}(\bar{x}; h) = 0,$$

where $d_A(x) := d(x, A)$ and $\psi_A(x) = 0$ if $x \in A$ and $+\infty$ otherwise.

The polar cone $N(A;\bar{x})$ of $T(A;\bar{x})$ is called the normal cone. We have $N(A;\bar{x})=\partial\psi_A(\bar{x})$, where for $f: E\to\mathbb{R}\cup\{-\infty,+\infty\}$ with $|f(\bar{x})|<\infty$ $\partial f(\bar{x})$ denotes the Clarke generalized gradient of f at $\bar{x}$ given by

$$\partial f(\bar{x})=\{y^*\in E^*: \langle y^*, h\rangle\leqq f^\uparrow(\bar{x};h)\forall hE\}.$$

COROLLARY 1.4. *Let M be any multifunction from E into F with $(\bar{x},\bar{y})\in GrM$. Then*:

(i) $d^0_{GrM}(\bar{x},\bar{y};h,k)\leqq\Delta^\uparrow_M(\bar{x},\bar{y};h,k)\leqq\psi^\uparrow_{GrM}(\bar{x},\bar{y};h,k)$;

(ii) $\partial d_{GrM}(\bar{x},\bar{y})\subset\partial\Delta_M(\bar{x},\bar{y})\subset N(GrM;\bar{x},\bar{y})$;

(iii) *If M is pseudo-Lipschitz at $(\bar{x},\bar{y})$ then $\Delta^\uparrow(\bar{x},\bar{y};h,k)=\Delta^0(\bar{x},\bar{y};h,k)\leqq 2K\,d^0_{GrM}(\bar{x},\bar{y};h,k)$, where K is the constant given by Proposition* 1.2.

Proof. We know (by what precedes with $M'(x,y)=GrM$, $E'=E\times F$ and $F'=E\times F$) that

$$d^0_{GrM}(\bar{x},\bar{y};h,k)=\limsup_{(x,y)\xrightarrow[t\downarrow 0]{M}(\bar{x},\bar{y})} t^{-1}\,d_{GrM}(x+th,y+tk),$$

and we easily see that

$$d_{GrM}(x,y)\leqq\Delta_M(x,y)\leqq\psi_{GrM}(x,y).$$

So the results are direct consequences of Propositions 1.3 and 1.2. □

We close this section by recalling a result of Clarke on distance functions that will be often used in the sequel of this paper.

PROPOSITION 1.5 (Clarke [7]). *Let $f: E\to\mathbb{R}$ be a function that is k-Lipschitz around $\bar{x}$. Assume that $\bar{x}$ gives a local minimum of f relative to a subset A of the metric space E. Then $\bar{x}$ gives a local unconstrained minimum of $f+kd_A$ (where $d_A(x)=d(x,A)$).*

2. Subdifferentials and singular subdifferentials.

Throughout this section E will be a *Banach space*.

The important notion of ε-Fréchet subgradient will be crucial in the sequel (see [20], [21], [34], and [35]). An element $x^*\in E^*$ is said to be an *F_ε-subgradient* to an extended real-valued function $f: E\to\bar{\mathbb{R}}:=\mathbb{R}\cup\{-\infty,+\infty\}$ at a point $\bar{x}$ where $|f(\bar{x})|<\infty$ if there exists a neighbourhood X of $\bar{x}$ such that

$$\langle x^*, x-\bar{x}\rangle\leqq f(x)-f(\bar{x})+\varepsilon\|x-\bar{x}\|\quad\text{for every } x\in X;$$

equivalently $\bar{x}$ gives a local minimum of the function

$$x\to f(x)-\langle x^*, x-\bar{x}\rangle+\varepsilon\|x-\bar{x}\|.$$

Let $\delta_\varepsilon f(\bar{x})$ the set of all F_ε-subgradients to f at $\bar{x}$.

The *subdifferential in the sense of Kruger-Mordukhovich* [20], [21] is the set $\delta f(\bar{x})$ of all $x^*\in E^*$ for which there exist $\varepsilon_n\downarrow 0$, $x_n\to\bar{x}$ with $f(x_n)\to f(\bar{x})$, $x^*_n\in\delta_{\varepsilon_n}f(\bar{x})$ with $x^*_n\xrightarrow{w^*}x^*$.

The singular subdifferential (see Treiman [34], [35]) associated with $\delta f(\bar{x})$ is the set $\delta^\infty f(\bar{x})$ of all $x^*\in F^*$ for which there exist $\varepsilon_n\downarrow 0$, $s_n\downarrow 0$, $x_n\to\bar{x}$ with $f(x_n)\to f(\bar{x})$, $x^*_n\in\delta_{\varepsilon_n}(s_nf)(x_n)$ with $x^*_n\xrightarrow{w^*}x^*$.

If A is a subset of E and $x\in A$, $\mathcal{N}(A,x)=\delta\Psi_A(x)$ is the Kruger-Mordukhovich normal cone to A at x.

The following deep result is due to Treiman [34], [35].

PROPOSITION 2.1. [34], [35]. *If E admits an equivalent norm that is Fréchet-differentiable off zero and f is lower semicontinuous, then*

$$\partial f(\bar{x})=\text{cl co}\,[\delta f(\bar{x})+\delta^\infty f(\bar{x})]$$

with the convention $\varnothing+A=\varnothing$.

Let us recall (Kruger [20]) a formula that gives an upper estimate of $\delta(f+g)$ in terms of δf and δg. Here we state Kruger's result in its less general form.

PROPOSITION 2.2. [20]. *Assume that $f: E \to \bar{\mathbb{R}}$ is lower semicontinuous with $|f(\bar{x})| < \infty$ and that $g: E \to \bar{\mathbb{R}}$ is locally Lipschitz around $\bar{x}$. If E admits an equivalent norm that is Fréchet differentiable off zero then*

$$\delta(f+g)(\bar{x}) \subset \delta f(\bar{x}) + \delta g(\bar{x}).$$

The following result will be needed in the search of estimates of subdifferentials of optimal value functions.

PROPOSITION 2.3. *Let f and g be two functions from E into $\bar{\mathbb{R}}$ with $|f(\bar{x})| < \infty$ and $|g(\bar{x})| < \infty$. If g is locally Lipschitz around $\bar{x}$, then*

$$\delta^{\infty}(f+g)(\bar{x}) = \delta^{\infty} f(\bar{x}).$$

Proof. Let $x^* \in \delta^{\infty}(f+g)(\bar{x})$. There exist $x_n \to \bar{x}$ with $(f+g)(x_n) \to (f+g)(\bar{x})$, $\varepsilon_n \downarrow 0$, $s_n \downarrow 0$ and $x_n^* \xrightarrow{w^*} x^*$ such that $x_n^* \in \delta_{\varepsilon_n}(s_n f + s_n g)(x_n)$. Consider an open neighbourhood X of $\bar{x}$ over which g is Lipschitz and denote by k a Lipschitz constant. We may assume that all $x_n \in X$. Choose for each $n \in \mathbb{N}$ a neighbourhood $X_n \subset X$ of x_n with

$$\langle x_n^*, x - x_n \rangle \leqq s_n(f+g)(x) - s_n(f+g)(x_n) + \varepsilon_n \|x - x_n\| \quad \text{for all } x \in X_n.$$

By Lipschitz continuity of g we have

$$\langle x_n^*, x - x_n \rangle \leqq s_n f(x) - s_n f(x_n) + \eta_n \|x - x_n\| \quad \text{for all } x \in X_n,$$

where $\eta_n := (\varepsilon_n + k s_n) \downarrow 0$. So $x_n^* \in \delta_{\eta_n} f(x_n)$ and $f(x_n) \to f(\bar{x})$ since $g(x_n) \to g(\bar{x})$ and this implies $x^* \in \delta^{\infty} f(\bar{x})$. Therefore we have

$$\delta^{\infty}(f+g)(\bar{x}) \subset \delta^{\infty} f(\bar{x}). \tag{2.1}$$

Writing $f = (f+g) + (-g)$ we also have by Lipschitz continuity of $(-g)$ and by relation (2.1),

$$\delta^{\infty} f(\bar{x}) \subset \delta^{\infty}(f+g)(\bar{x}),$$

and the proof is complete. □

COROLLARY 2.4. *Let $g: E \to \bar{\mathbb{R}}$ be locally Lipschitz around $\bar{x}$. Then*

$$\delta^{\infty} g(\bar{x}) = \{0\}.$$

Proof. It suffices to write $g = 0 + g$. □

In the same vein we have the following result.

PROPOSITION 2.5. *Let f and g be two functions from E into $\bar{\mathbb{R}}$ with $|f(\bar{x})| < \infty$ and $|g(\bar{x})| < \infty$. Assume that g is locally Lipschitz around $\bar{x}$. Then if $x_n \to \bar{x}$ with $f(x_n) \to f(\bar{x})$, $\varepsilon_n \downarrow 0$, $s_n \downarrow 0$, $x_n^* \xrightarrow{w^*} x^*$, $x_n^* \in \delta_{\varepsilon_n}(f + s_n g)(x_n)$, then*

$$x^* \in \delta f(\bar{x}).$$

Proof. For x in some neighbourhood X_n of x_n we have

$$\begin{aligned} \langle x_n^*, x - x_n \rangle &\leqq f(x) - f(x_n) + s_n(g(x) - g(x_n)) + \varepsilon_n \|x - x_n\| \\ &\leqq f(x) - f(x_n) + \eta_n \|x - x_n\|, \end{aligned}$$

where $\eta_n := (\varepsilon_n + k s_n) \downarrow 0$ and k is a Lipschitz constant for g around the point $\bar{x}$. Then $x_n^* \in \delta_{\eta_n} f(x_n)$ and hence $x^* \in \delta f(\bar{x})$. □

We need to consider the *Clarke singular generalized gradient* $\partial^\infty f(\bar{x})$, which is the set of all $x^* \in E^*$ for which there exist $s_n \downarrow 0$, $x_n \to x$ with $f(x_n) \to f(\bar{x})$, $x_n^* \in s_n\, \partial f(x_n)$ with $x_n^* \xrightarrow{w^*} x^*$. (Recall that the set of points $(x, f(x))$ where $\partial f(x) \neq \varnothing$ is dense in $Gr(f)$ whenever f is lower semicontinuous; see McLinden [22].)

PROPOSITION 2.6. $\partial^\infty f(\bar{x}) \subset \partial^\infty f(\bar{x})$.

Proof. Let $x^* \in \delta^\infty f(\bar{x})$. There exist $\varepsilon_n \downarrow 0$, $s_n \downarrow 0$, $x_n \to \bar{x}$ with $f(x_n) \to f(\bar{x})$, $x_n^* \in \delta_{\varepsilon_n} s_n f(x_n)$ with $x_n^* \xrightarrow{w^*} x^*$. Using the definition of $\delta_{\varepsilon_n} f(x_n)$ it is not difficult to see that

$$\langle x_n^*, h\rangle \leqq s_n f^\uparrow(x_n; h) + \varepsilon_n \|h\| \quad \text{for all } h \in E,$$

and hence by subdifferential calculus (see [27]) we have

$$x_n^* \in s_n\, \partial f(x_n) + \varepsilon_n B^*,$$

where B^* is the closed unit ball with center zero of E^*. So we can conclude that $x^* \in \partial^\infty f(\bar{x})$. □

Let us show that Δ_M also allows us to describe $\mathcal{N}(M; \bar{x}, \bar{y})$.

PROPOSITION 2.7. *Let F be a Banach space, $M: E \rightrightarrows F$ a multifunction with closed graph, and $(\bar{x}, \bar{y}) \in GrM$. Then*

$$\mathcal{N}(M; \bar{x}, \bar{y}) = \bigcup_{s \geqq 0} s\delta\Delta_M(\bar{x}, \bar{y}).$$

Proof. (i) Let $(x^*, y^*) \in \delta\Delta_M(\bar{x}, \bar{y})$. There exist $(x_n^*, y_n^*) \xrightarrow{w^*} (x^*, y^*)$, $(x_n, y_n) \xrightarrow{\Delta_M} (\bar{x}, \bar{y})$, $\varepsilon_n \downarrow 0$ and $\rho_n > 0$ such that

$$\langle x_n^*, v\rangle + \langle y_n^*, w\rangle \leqq \Delta_M(x_n + v, y_n + w) - \Delta_M(x_n, y_n) + \varepsilon_n(\|v\| + \|w\|)$$

for each (v, w) with $\|v\| + \|w\| \leqq \rho_n$. We may suppose $y_n \notin M(x_n)$ since otherwise the proof is finished. Choose $r_n > 0$ with $r_n^2 < \min(\rho_n^2, (1/n)\Delta_M(x_n, y_n))$. There exist $s_n \downarrow 1$ such that $(s_n - 1)\Delta_M(x_n, y_n) < r_n^2$. If we choose $y_n' \in M(x_n)$ satisfying $\|y_n - y_n'\| \leqq s_n \Delta_M(x_n, y_n)$ we get

$$\begin{aligned}\langle x_n^*, v\rangle + \langle y_n^*, w\rangle &\leqq \Delta_M(x_n + v, y_n + w) - s_n^{-1}\|y_n - y_n'\| + \varepsilon_n(\|v\| + \|w\|)\\ &\leqq \Delta_M(x_n + v, y_n' + w) + (1 - s_n^{-1})\|y_n - y_n'\| + \varepsilon_n(\|v\| + \|w\|)\end{aligned}$$

for all v, w with $\|v\| + \|w\| \leqq \rho_n$. Then for all $(x, y) \in M \cap B((x_n, y_n'), \rho_n)$ we have

$$\langle x_n^*, x - x_n\rangle + \langle y_n^*, y - y_n'\rangle \leqq (1 - s_n^{-1})\|y_n - y_n'\| + \varepsilon_n(\|x - x_n\| + \|y - y_n'\|),$$

and hence for $\gamma_n^2 := (1 - s_n^{-1})\|y_n - y_n'\|$ we get

$$0 \leqq \langle -x_n^*, x - x_n\rangle + \langle -y_n^*, y - y_n'\rangle + \varepsilon_n(\|x - x_n\| + \|y - y_n'\|) + \gamma_n^2.$$

By the Ekeland variational principle [2] there exist $(x_n'', y_n'') \in M \cap B((x_n, y_n'), \rho_n)$ satisfying $\|x_n'' - x_n\| + \|y_n'' - y_n'\| \leqq \gamma_n$ and such that for $(x, y) \in M \cap B((x_n, y_n'), \rho_n)$

$$\begin{aligned}\langle -x_n^*, x_n'' - x_n\rangle + \langle -y_n^*, y_n'' - y_n'\rangle + \varepsilon_n(\|x_n'' - x_n\| + \|y_n'' - y_n'\|) &\leqq \langle -x_n^*, x - x_n\rangle\\ &\quad + \langle -y_n^*, y - y_n'\rangle + \varepsilon_n(\|x - x_n\| + \|y - y_n'\|)\\ &\quad + \gamma_n(\|x - x_n''\| + \|y - y_n''\|),\end{aligned}$$

and hence

$$\langle x_n^*, x - x_n''\rangle + \langle y_n^*, y - y_n''\rangle \leqq (\varepsilon_n + \gamma_n)(\|x - x_n''\| + \|y - y_n''\|). \tag{2.2}$$

As $\gamma_n^2 = (1 - s_n^{-1})\|y_n - y_n'\| \leqq s_n(1 - s_n^{-1})\Delta_M(x_n, y_n) < r_n^2$ we have $\alpha_n := r_n - \gamma_n > 0$. Moreover,

$$\|x - x_n''\| + \|y - y_n''\| \leqq \alpha_n \Rightarrow \|x - x_n\| + \|y - y_n'\| \leqq \|x - x_n''\| + \|y - y_n''\| + \gamma_n \leqq r_n \leqq \rho_n.$$

Therefore for all $(x,y)\in M\cap B((x''_n,y''_n),\alpha_n)$ we have by (2.2)

$$\langle x^*_n, x-x''_n\rangle+\langle y^*_n, y-y''_n\rangle\leqq(\varepsilon_n+\gamma_n)(\|x-x''_n\|+\|y-y''_n\|),$$

and hence $(x^*_n,y^*_n)\in\delta_{\varepsilon_n+\gamma_n}\psi_M(x''_n,y''_n)$. So $(x^*,y^*)\in\delta\psi_M(\bar{x},\bar{y})=\mathcal{N}(M;\bar{x},\bar{y})$.

(ii) Consider now $(x^*,y^*)\in\mathcal{N}(M;\bar{x},\bar{y})$. Choose $(x^*_n,y^*_n)\xrightarrow{w^*}(x^*,y^*)$, $(x_n,y_n)\xrightarrow{M}(\bar{x},\bar{y})$, $\varepsilon_n\downarrow 0$ and $r_n>0$ such that

$$\langle x^*_n, x-x_n\rangle+\langle y^*_n, y-y_n\rangle\leqq\varepsilon_n(\|x-x_n\|+\|y-y_n\|)$$

for every $(x,y)\in M\cap B((x_n,y_n),3r_n)$. Choose $k>0$ with $\|x^*_n\|+\|y^*_n\|\leqq k$ and $\varepsilon_n\leqq k$ for all $n\in\mathbb{N}$. By Proposition 1.5 for $B_n:=B((x_n,y_n),3r_n)$ we have

$$\langle x^*_n, x-x_n\rangle+\langle y^*_n, y-y_n\rangle\leqq 2kd((x,y),M\cap B_n)+\varepsilon_n(\|x-x_n\|+\|y-y_n\|)$$

for every $(x,y)\in B_n$ and hence

$$\langle x^*_n, x-x_n\rangle+\langle y^*_n, y-y_n\rangle\leqq 2k\Delta_M(x,y)+\varepsilon_n(\|x-x_n\|+\|y-y_n\|)$$

for every $(x,y)\in B((x_n,y_n),r_n)$. Therefore $(x^*_n,y^*_n)\in 2k\delta_{\varepsilon' n}\Delta_M(x_n,y_n)$ with $\varepsilon'_n=(2k)^{-1}\varepsilon_n$, which ensures that $(x^*,y^*)\in 2k\delta\Delta_M(\bar{x},\bar{y})$. So the proof is complete. □

3. Subdifferentials of optimal value functions. Let $f:E\times F\to\bar{\mathbb{R}}$ be lower semicontinuous and let C be a closed subset of $E\times F$ where E and F are two Banach spaces. We consider the perturbed problem $\mathcal{P}_u$, which consists in minimizing for u the function $f(\cdot,u)$ over all x satisfying $(x,u)\in C$, and we define the function $m:F\to\bar{\mathbb{R}}$ and the multifunction $S:F\rightrightarrows E$ by

$$m(u)=\inf_{x\in C_u}f(x,u)\quad\text{and}\quad S(u)=\{x\in C_u : m(u)=f(x,u\},$$

where $C_u=\{x\in E\colon (x,u)\in C\}$.

We say that the family of problems $\mathcal{P}_u$ satisfies *condition* (K) at $\bar{u}$ if m is lower semicontinuous around $\bar{u}$ with $m(\bar{u})$ finite, S takes nonempty values on some neighbourhood of $\bar{u}$, and for every sequence $(\varepsilon_n,u_n)\to(0_+,\bar{u})$ with $m(u_n)\to m(\bar{u})$ and $\delta_{\varepsilon_n}(u_n)\neq\emptyset$ there exist some sequence $(x_n)_n$ admitting a cluster point and such that $x_n\in S(u_n)$ for n sufficiently large.

Note that such a cluster point necessarily belongs to $S(\bar{u})$.

General assumptions ensuring condition (K) can be found in [7], [13], [29], and [31]. Condition (K) is, for example, satisfied whenever the projection of C on E is compact or whenever the multifunction S admits a local selection that is continuous at $\bar{u}$. It also holds if there is a mapping $h:F\to E$ continuous at $\bar{u}$ and a compact $H\subset E$ such that $C_u\subset h(u)+H$ for u near $\bar{u}$.

All the results in the rest of the paper will justify, each in its own situation, the intuitively clear fact: for f convex in both variables $u^*\in\partial m(\bar{u})$, $m(u)=\inf_x f(x,u)$ and $m(\bar{u})=f(\bar{x},\bar{u})$ we have $(0,u^*)\in\partial f(\bar{x},\bar{u})$.

In the same vein it is also clear that for $u^*\in\delta_\varepsilon m(\bar{u})$ we have $(0,u^*)\in\delta_\varepsilon f(\bar{x},\bar{u})$. This illustrates how the Kruger–Mordukhovich subdifferential can be useful in obtaining estimates of subdifferentials of optimal value functions.

Our aim in this section and the following ones is to give estimates for the Clarke subdifferential of m. In light of the important result of Treiman recalled in Proposition 2.1 it will be enough to establish estimates for $\delta m(\bar{u})$ and $\delta^\infty m(\bar{u})$.

Throughout this section (because of Propositions 2.1 and 2.2) we assume that E and F admit *equivalent norms that are Fréchet-differentiable* off zero. By [15] the closed balls of E^* and F^* are w^*-sequentially compact.

PROPOSITION 3.1. *Let $f: E \times F \to \mathbb{R}$ be locally Lipschitz, let $M: E \rightrightarrows F$ be a multifunction with closed graph, and let A be a closed subset of E. Let*

$$m(u) = \inf_x f(x, u)\colon u \in M(x), x \in A\}.$$

Assume that condition (K) is satisfied at $\bar{u}$ and that M is pseudo-Lipschitz at each point in $S(\bar{u}) \times \{\bar{u}\}$. Then:

(i) *For each $u^* \in \delta m(\bar{u})$ there exist $\bar{x} \in S(\bar{u})$ and $\lambda > 0$ such that $(0, u^*) \in \delta(f + \lambda\Delta_M + \lambda d_{A\times F})(\bar{x}, \bar{u}) \subset \delta f(\bar{x}, \bar{u}) + \lambda\delta\Delta_M(\bar{x}, \bar{u}) + \lambda\delta d_A(\bar{x}) \times \{0\}$;*

(ii) *For each $u^* \in \delta^\infty m(\bar{u})$ there exist $\bar{x} \in S(\bar{u})$ and $\gamma > 0$ such that $(0, u^*) \in \gamma\delta(\Delta_M + d_{A\times F})(\bar{x}, \bar{u}) \subset \gamma\delta\Delta_M(\bar{x}, \bar{u}) + \gamma\delta_A(\bar{x}) \times \{0\}$.*

Proof. (i) Let $u^* \in \delta m(\bar{u})$. There exist $\varepsilon_n \downarrow 0$, $u_n \to \bar{u}$ with $m(u_n) \to m(\bar{u})$, $u_n^* \xrightarrow{w^*} u^*$ such that the function

$$u \to m(u) - \langle u_n^*, u - u_n\rangle + \varepsilon_n \|u - u_n\|$$

attains a local minimum at u_n. By assumption (K) we may suppose that there exists $x_n \in S(u_n)$ with (x_n) converging to some $\bar{x} \in S(\bar{u})$. Then if we consider the multifunction T defined by $T(x) = M(x)$ if $x \in A$ and $T(x) = \emptyset$ otherwise, we have for every $(x, u) \in GrT$ and u near u_n

$$\begin{aligned} -\langle u_n^*, u_n\rangle + f(x_n, u_n) &\leqq \varepsilon_n \|u - u_n\| - \langle u_n^*, u\rangle + m(u) \\ &\leqq \varepsilon_n \|u - u_n\| - \langle u_n^*, u\rangle + f(x, u). \end{aligned}$$

Therefore for n sufficiently large, by Proposition 1.5, (x_n, u_n) gives a local minimum of the function

$$(x, u) \to \varepsilon_n \|u - u_n\| - \langle u_n^*, u\rangle + f(x, u) + (\varepsilon_n + k + \|u_n^*\|) d(x, u; GrT),$$

where k is a Lipschitz constant for f around $(\bar{x}, \bar{u})$ and hence (x_n, u_n) also gives a local minimum relative to $A \times F$ for the function

$$(x, u) \to \varepsilon_n \|u - u_n\| - \langle u_n^*, u\rangle + f(x, u) + (\varepsilon_n + k + \|u_n^*\|)\Delta_M(x, u).$$

Denoting by l a pseudo-Lipschitz constant for M around $(\bar{x}, \bar{u})$ we obtain that (x_n, u_n) gives a local minimum of the function

$$\begin{aligned} (x, u) \to \varepsilon_n \|u - u_n\| - \langle u_n^*, u\rangle + f(x, u) + (\varepsilon_n + \|u_n^*\| + k)\Delta_M(x, u) \\ + 3(1 + l)(\varepsilon_n + \|u_n^*\| + k) d_A(x). \end{aligned}$$

Choose a real number $\lambda \geqq 3(1 + l) \sup_n (\varepsilon_n + \|u_n^*\| + k)$. Then

$$(0, u_n^*) \in \delta_{\varepsilon_n}(f + \lambda\Delta_M + \lambda d_{A\times F})(x_n, u_n) \tag{3.1}$$

and, as $(f + \lambda\Delta_M + \lambda d_{A\times F})(x_n, u_n) \to (f + \lambda\Delta_M + \lambda d_{A\times F})(\bar{x}, \bar{u})$, this yields

$$(0, u^*) \in \delta(f + \lambda\Delta_M + \lambda d_{A\times F})(\bar{x}, \bar{u}) \subset \delta f(\bar{x}, \bar{u}) + \lambda\delta\Delta_M(\bar{x}, \bar{u}) + \lambda\delta d_A(\bar{x}) \times \{0\}.$$

(ii) As in part (i) for some $s_n \downarrow 0$ and $\gamma \geqq 3(1 + l) \sup_n (\varepsilon_n + \|u_n^*\| + s_n k)$ the function

$$(x, u) \to \varepsilon_n \|u - u_n\| - \langle u_n^*, u\rangle + s_n f(x, u) + \gamma\Delta_M(x, u) + \gamma d_{A\times F}(x, u)$$

attains a local minimum at (x_n, u_n) and hence

$$(0, u_n^*) \in \delta_{\varepsilon_n}(s_n f + \gamma\Delta_M + \gamma d_{A\times F})(x_n, u_n).$$

Proposition 2.5 ensures that

$$(0, u^*) \in \gamma\delta(\Delta_M + d_{A\times F})(\bar{x}, \bar{u}) \subset \gamma\delta\Delta_M(\bar{x}, \bar{u}) + \gamma\delta d_A(\bar{x}) \times \{0\},$$

which completes the proof. □

Remark. By (3.1) we see that it is enough to require instead of condition (K) the convergence of $(x_n)_n$ to some $\bar{x}$ with respect to some topology for which we can pass to the limit in (3.1). This condition holds in the perturbed control problems considered by Clarke [7] and Clarke and Loewen [8]. Also, for

$$m(u)=\inf_x\{(f(Jx)\colon u\in M(x), J_0(x)\in A_0\}$$

whenever M is a convex multifunction, A_0 is a subset of some finite-dimensional space X_0, J and J_0 are two continuous surjective linear mappings from X into the finite-dimensional spaces X' and X_0, respectively, and the projection of $\{(x, u)\colon u\in M(x), J_0(x)\in A_0\}$ on X is weakly compact.

If M is a multifunction from E into F with $(\bar{x}, \bar{y})\in GrM$, we define the Kruger-Mordukhovich co-derivative of M at (x, y) as the multifunction from F^* into E^* satisfying

$$x^*\in D^*M(\bar{x}, \bar{y})(y^*)\Leftrightarrow(x^*, -y^*)\in\mathcal{N}(GrM; \bar{x}, \bar{y}),$$

where $\mathcal{N}(GrM; \bar{x}, \bar{y})$ denotes the Kruger-Mordukhovich normal cone to GrM.

COROLLARY 3.2. *Let $f\colon E\to\mathbb{R}$ be locally Lipschitz, let $M\colon E\rightrightarrows F$ be a multifunction with closed graph, and let A be a closed subset of E. Let*

$$m(u)=\inf_x\{f(x)\colon u\in M(x), x\in A\}.$$

Assume that condition (K) is satisfied at $\bar{u}$ and that M is pseudo-Lipschitz at each point in $S(\bar{u})\times\{\bar{u}\}$. Then

$$\delta m(\bar{u})\subset\bigcup_{\bar{x}\in S(\bar{u})}\{u^*\in F^*\colon 0\in\partial f(x)+D^*M(\bar{x})(-u^*)+N(A; \bar{x})\}$$

and

$$\delta^\infty m(\bar{u})\subset\bigcup_{\bar{x}\in S(\bar{u})}\{u^*\in F^*\colon 0\in D^*M(\bar{x})(-u^*)+N(A; \bar{x})\}.$$

Proof. Let $u^*\in\delta m(\bar{u})$. Then by Proposition 3.1 there exist $\bar{x}\in S(\bar{u})$, $\lambda>0$ such that

$$(0, u^*)\in\delta f(\bar{x})\times\{0\}+\lambda\delta\Delta_M(\bar{x}, \bar{u})+\lambda\delta d_\Delta(\bar{x})\times\{0\},$$

and hence by Proposition 2.7 there are

$$x_1^*\in\delta f(\bar{x}),\quad (x_2^*, u_2^*)\in\mathcal{N}(M; \bar{x}, \bar{u}),\quad x_3^*\in N(A; \bar{x})$$

satisfying

$$x_1^*+x_2^*+x_3^*=0\quad\text{and}\quad u^*=u_2^*.$$

Therefore

$$0\in\partial f(\bar{x})+D^*M(\bar{x})(-u^*)+N(A; \bar{x}),$$

and this proves the first inclusion. The proof of the second inclusion is similar. □

Before stating the next corollary let us give another description of D^*G when G is a locally Lipschitz single-valued mapping.

PROPOSITION 3.3. *Let $G\colon E\to F$ be a mapping that is k-Lipschitz around $\bar{x}$. For $\bar{y}:=G(\bar{x})$ we have $D^*G(\bar{x}, \bar{y})(y^*)=\hat{\delta}(y^*\circ G)(\bar{x})$, where $\hat{\delta}(y^*\circ G)(\bar{x})$ is the set of $x^*\in E^*$ for which there exist $y_n^*\xrightarrow{w^*}y^*$, $\varepsilon_n\downarrow 0$, $x_n\to\bar{x}$, $x_n^*\xrightarrow{w^*}x^*$ satisfying $x_n^*\in\delta_{\varepsilon_n}(y_n^*\circ G)(x_n)$.*

Proof. Let $(x^*, -y^*) \in \mathcal{N}(GrG; \bar{x}, \bar{y})$. There exist $x_n^* \xrightarrow{w^*} x^*$, $y_n^* \xrightarrow{w^*} y^*$, $\varepsilon_n \downarrow 0$, $x_n \to \bar{x}$, and $r_n > 0$ such that for any $x \in B(x_n, r_n)$

$$\langle x_n^*, x - x_n\rangle - \langle y_n^*, G(x) - G(x_n)\rangle \leqq \varepsilon_n(\|x - x_n\| + \|G(x) - G(x_n)\|) \leqq \varepsilon_n(1+k)\|x - x_n\|,$$

and hence $x_n^* \in \partial_{\varepsilon_n'}(y_n^* \circ G)(x_n)$ for $\varepsilon_n' = \varepsilon_n(1+k)$. So $x^* \in \hat{\partial}(y^* \circ G)(\bar{x})$.

The converse implication is similar. □

PROPOSITION 3.4 (Kruger [20]). *Let E and F be two Banach spaces that admit equivalent Fréchet differentiable (off zero) norms. Let $G: E \to F$ and $g: F \to \mathbb{R}$ be two locally Lipschitz mappings and $\bar{y} := G(\bar{x})$.*

Then $\partial(g \circ G)(\bar{x}) \subset \bigcup \{\hat{\partial}(y^* \circ G)(\bar{x}): y^* \in \partial g(\bar{y})\}$.

Remark. $\hat{\partial}(y^* \circ G)(\bar{x}) = \partial(y^* \circ G)(\bar{x})$ whenever F is finite-dimensional.

COROLLARY 3.5. *Let $f: E \to \mathbb{R}$ and $G: E \to F$ be two locally Lipschitz mappings, and let A and B be two closed subsets in E and F, respectively. Let*

$$m(u) = \inf_x \{f(x): G(x) + u \in B, x \in A\}.$$

Assume that condition (K) is satisfied at $\bar{u}$. Then

(i) $$\partial m(\bar{u}) \subset \bigcup_{\bar{x} \in S(\bar{u})} \{u^* \in \mathcal{N}(B; G(\bar{x}) + \bar{u}): 0 \in \partial f(\bar{x}) + \hat{\partial}(u^* \circ G)(\bar{x}) + \mathcal{N}(A; \bar{x})\}$$

and

$$\partial^\infty m(\bar{u}) \subset \bigcup_{\bar{x} \in S(\bar{u})} \{u^* \in \mathcal{N}(B; G(\bar{x}) + \bar{u}): 0 \in \hat{\partial}(u^* \circ G)(\bar{x}) + \mathcal{N}(A; \bar{x})\}.$$

(ii) *If F is finite-dimensional, both inclusions above hold with $\partial(u^* \circ G)(\bar{x})$ replacing $\hat{\partial}(u^* \circ G)(\bar{x})$.*

Proof. Put $M(x) = -G(x) + B$. Then M is locally Lipschitz and

$$\Delta_M(x, u) = d(u, -G(x) + B) = d_B \circ H(x, u),$$

where $H(x, u) = G(x) + u$. Let $u^* \in \partial m(\bar{u})$. By Proposition 3.1 there exist $\bar{x} \in S(\bar{u})$, $\lambda > 0$ such that

$$(0, u^*) \in \partial f(\bar{x}) \times \{0\} + \lambda \partial \Delta_M(\bar{x}, \bar{u}) + \lambda \partial d_A(\bar{x}) \times \{0\}$$

and hence putting $\bar{z} = G(\bar{x}) + \bar{u}$ we obtain by Proposition 3.4 that there are $x_1^* \in \partial f(\bar{x})$, $y^* \in \lambda \partial d_B(\bar{z}) \subset \mathcal{N}(B; \bar{z})$, $(x_2^*, u_2^*) \in \hat{\partial}(y^* \circ G)(\bar{x}) \times \{y^*\}$, $x_3^* \in \lambda \partial d_A(\bar{x}) \subset \mathcal{N}(A; \bar{x})$ satisfying

$$x_1^* + x_2^* + x_3^* = 0, \qquad u_2^* = y^* = u^*.$$

Therefore $u^* \in \mathcal{N}(B; \bar{z})$ and

$$0 \in \partial f(\bar{x}) + \hat{\partial}(u^* \circ G)(\bar{x}) + \mathcal{N}(A; \bar{x}).$$

The first inclusion (i) is then proved and the proof of the second one is the same. So the proof is complete since (ii) is a direct consequence of (i) for $\hat{\partial}(u^* \circ G)(\bar{x}) = \partial(u^* \circ G)(\bar{x})$ whenever F is finite-dimensional. □

4. Perturbed problems with metrically regular constraints. In this section we consider the perturbed problem $\mathcal{P}_u$

$$m(u) = \inf_x \{f(x, u): 0 \in M(x, u), (x, u) \in A\},$$

where A is a closed subset in $E \times F$, $M: E \times F \rightrightarrows G$ is a multifunction of closed graph from $E \times F$ into a Banach space G.

Let us recall that M is *metrically regular* at $(\bar{x}, \bar{u}, 0) \in GrM$ with respect to A if there exist $\gamma > 0$, $X \times U$ a neighbourhood of $(\bar{x}, \bar{u})$ such that

$$(4.1) \qquad d((x, u), A \cap M^{-1}(0)) \leqq \gamma d(0, M(x, u)) \quad \text{for all } (x, u) \in (X \times U) \cap A.$$

Actually, many results on verifiable conditions ensuring metric regularity are known (see [1], [16], [17], [19], [25], [26], and the references therein).

LEMMA 4.1. *Assume that f is locally Lipschitz. Let $u^* \in \partial_\varepsilon m(\bar{u})$, $\bar{x} \in S(\bar{u})$, and $\alpha > (\varepsilon + \|u^*\| + k)$ (where k is a Lipschitz constant for f around $(\bar{x}, \bar{u})$). If M is metrically regular at $(\bar{x}, \bar{u}, 0)$, then the point $(\bar{x}, \bar{u})$ is a local minimum for the function*

$$(x, u) \to \varepsilon \|u - \bar{u}\| - \langle u^*, u\rangle + f(x, u) + \alpha\gamma d(0, M(x, u)) + q(x, u),$$

where $q := \alpha(1 + \gamma l)d_A$ if M is l-pseudo-Lipschitz at $(\bar{x}, \bar{u}, 0)$ and ψ_A otherwise.

Proof. By assumption there exists a neighbourhood U_0 of $\bar{u}$ over which $\bar{u}$ gives a minimum of the function

$$u \to \varepsilon \|u - \bar{u}\| - \langle u^*, u\rangle + m(u).$$

Choose a neighbourhood $X \times U$ of $(\bar{x}, \bar{u})$ with $U \subset U_0$ and over which (4.1) is satisfied. Then for $(x, u) \in (X \times U) \cap A \cap M^{-1}(0)$ we have

$$\begin{aligned} -\langle u^*, \bar{u}\rangle + f(\bar{x}, \bar{u}) &\leqq \varepsilon \|u - \bar{u}\| - \langle u^*, u\rangle + m(u) \\ &\leqq \varepsilon \|u - \bar{u}\| - \langle u^*, u\rangle + f(x, u), \end{aligned}$$

and hence, by Proposition 1.5, $(\bar{x}, \bar{u})$ gives a local minimum of the function

$$(x, u) \to \varepsilon \|u - \bar{u}\| - \langle u^*, u\rangle + f(x, u) + \alpha d((x, u), A \cap M^{-1}(0)).$$

By metrical regularity $(\bar{x}, \bar{u})$ gives a local minimum relative to A of the function

$$(x, u) \to \varepsilon \|u - \bar{u}\| - \langle u^*, u\rangle + f(x, u) + \alpha\gamma d(0, M(x, u)),$$

and hence, by Proposition 1.5 again, $(\bar{x}, \bar{u})$ gives a local minimum of the function

$$(x, u) \to \varepsilon \|u - \bar{u}\| - \langle u^*, u\rangle + f(x, u) + \alpha\gamma d(0, M(x, u)) + q(x, u). \qquad \square$$

We denote by p_M the function defined by

$$p_M(x, u) = d(0, M(x, u))$$

and we assume in the rest of this section that E and F admit equivalent *Fréchet differentiable norms.*

PROPOSITION 4.2. *Assume f is locally Lipschitz, M is regular with respect to A at each point in $S(\bar{u}) \times \{\bar{u}\} \times \{0\}$, and condition (K) is satisfied at $\bar{u}$. Let $u^* \in \delta m(\bar{u})$:*

(i) *There exist $\bar{x} \in S(\bar{u})$ and $\alpha > 0$ such that*

$$(0, u^*) \in \delta(f + \alpha\gamma p_M + \psi_A)(\bar{x}, \bar{u}).$$

(ii) *If in addition M is pseudo-Lipschitz at each point in $S(\bar{u}) \times \{\bar{u}\} \times \{0\}$, then there exist $\bar{x} \in S(\bar{u})$, $\alpha > 0$, and $l > 0$ such that*

$$(0, u^*) \in \delta(f + \alpha\gamma p_M + \alpha(1 + \gamma l)d_A)(\bar{x}, \bar{u}).$$

Proof. Choose $u_n \to \bar{u}$ with $m(u_n) \to m(\bar{u})$, $\varepsilon_n \downarrow 0$, $u_n^* \xrightarrow{w^*} u^*$ with $u_n^* \in \partial_{\varepsilon_n} m(u_n)$ and choose also $x_n \in S(u_n)$. By condition (K) we may suppose $x_n \to \bar{x} \in S(\bar{u})$. If we denote by k a Lipschitz constant for f around $(\bar{x}, \bar{u})$ and we choose $\alpha > (1 + \|u^*\| + k)$, Lemma 4.1 ensures that (for n sufficiently large) the point (x_n, u_n) gives a local minimum of the function

$$(x, u) \to \varepsilon_n \|u - u_n\| - \langle u_n^*, u\rangle + f(x, u) + \alpha\gamma p_M(x, u) + \psi_A(x, u).$$

As

$$(f+\alpha\gamma p_M+\psi_A)(x_n, u_n)\to(f+\alpha\gamma p_M+\psi_A)(\bar{x}, \bar{u})$$

we obtain

$$(0, u^*)\in\delta(f+\alpha\gamma+\psi_A)(\bar{x}, \bar{u})$$

and this proves (i). The proof of (ii) is exactly the same. □

PROPOSITION 4.3. *Assume f is locally Lipschitz, M is regular with respect to A at each point in $S(\bar{u})\times\{\bar{u}\}\times\{0\}$, and condition (K) is satisfied at $\bar{u}$. Let $u^*\in\delta^\infty(\bar{u})$:*

(i) *There exist $\bar{x}\in S(\bar{u})$ and $\alpha>0$ such that*

$$(0, u^*)\in\delta(\alpha\gamma p_M+\psi_A)(\bar{x}, \bar{u}).$$

(ii) *If in addition M is pseudo-Lipschitz at each point in $S(\bar{u})\times\{\bar{u}\}\times\{0\}$, then there exist $\bar{x}\in S(\bar{u})$, $\alpha>0$, and $l>0$ such that*

$$(0, u^*)\in\delta(\alpha\gamma p_M+\alpha(1+\gamma l)d_A)(\bar{x}, \bar{u}).$$

Proof. The proof is similar to that of Proposition 4.2. □

5. Estimates in terms of subdifferentials of support functions. In this section M is still a multifunction from a Banach space E into a Banach space F. For $x\in E$ and $y^*\in F^*$ we put

$$h(x, y^*)=\sup\{\langle y^*, y\rangle: y\in M(x)\}.$$

In all the sequel we assume that $M(x)$ is a *closed convex* subset in F for each $x\in E$ and that F is *reflexive.* So there exists an equivalent norm on F that is Fréchet differentiable off zero and throughout this section F will always be endowed with *such a norm.* Moreover, by the lopsided minimax theorem (see [2]) or by the Hahn–Banach theorem, it is not difficult to see that whenever $M(x)\neq\emptyset$

$$\Delta_M(x, y)=\max_{y^*\in B_{F^*}}(\langle y^*, y\rangle-h(x, y^*)), \tag{5.1}$$

where B_{F^*} denotes the closed unit ball centered at the origin of F^*.

In order to give a relationship between $\partial\Delta$ and ∂h let us begin with the following result, which is a variant of Theorem 2.8.2 of Clarke [6].

PROPOSITION 5.1. *Let G be a Banach space with B_{G^*} w^*-sequentially compact, and let S be a sequentially compact space and $\bar{x}\in G$. Let $g_s: G\to\bar{\mathbb{R}}$ for each $s\in S$. Assume that there exists a neighbourhood X of $\bar{x}$ in G and a real number $k\geqq 0$ such that for each $s\in S$ the function g_s is finite on X and k-Lipschitz on X. Assume also that the supremum in*

$$g(x):=\sup\{g_s(x): s\in S\}$$

is attained for each $x\in X$. Then g is finite and k-Lipschitz on X and

$$\partial g(\bar{x})\subset\operatorname{cl}\operatorname{co}\bigcup\{\hat{\partial}g_s(\bar{x}): s\in S(\bar{x})\},$$

where $S(\bar{x})=\{s\in S: g(\bar{x})=g_s(\bar{x})\}$ and

$$\hat{\partial}g_s(\bar{x})=\{\lim x_n^*: x_n^*\in\partial g_{s_n}(x_n), s_n\to s, x_n\to\bar{x}'\}.$$

Proof. It suffices to repeat (with very slight modification) the arguments of the first part of the proof of Theorem 2.8.2 in Clarke [7]. □

Remarks. (1) If G admits an equivalent Gateaux-differentiable norm, then (see [15]) B_{G^*} is w^*-sequentially compact.

(2) If $g_s^\Box(\bar{x}; \bar{v})$ denotes the following "lim sup" of the difference quotient

$$g_s^\Box(x; v) = \limsup_{\substack{(x,r)\to(\bar{x},s)\\ t\downarrow 0}} t^{-1}[g_r(x+tv) - g_r(x)],$$

then $g_s^\Box(\bar{x}; \cdot)$ is k-Lipschitz on G and hence the set

$$\partial^\Box g_s(\bar{x}) := \{x^* \in G^*: \langle x^*, v\rangle \leqq g_s^\Box(\bar{x}; v) \forall v \in G\}$$

is nonempty, w^*-compact and convex, and moreover it is not difficult to verify that

$$\hat{\partial} g_s(\bar{x}) \subset \partial^\Box g_s(\bar{x})$$

and that the relations

$$x_n^* \in \partial^\Box g_{s_n}(x_n), \quad x_n \to \bar{x}, \quad s_n \to s, \quad x_n^* \xrightarrow{w^*} x^*$$

ensure that $x^* \in \partial^\Box g_s(\bar{x})$.

Furthermore $\partial^\Box g_s(\bar{x}) = \operatorname{cl} \operatorname{co} \hat{\partial} g_s(\bar{x})$ whenever S is metrizable and G is separable.

Let $r > 0$, $M_r(x) = M(x) + rB_F$, and $h_r(x, y^*) = \sup \{\langle y^*, y\rangle: y \in M_r(x)\} = h(x, y^*) + r\|y^*\|_*$. As the dual norm of F^* is strictly convex (since the norm of F is Gateaux-differentiable off zero; see [12]), then for each $x \in E$ the function $y^* \to h_r(x, y^*)$ is strictly convex on its domain.

If M is locally Lipschitz around $\bar{x}$ in the sense that there exists $l > 0$ and a neighbourhood X of $\bar{x}$ in E such that

$$M(x') \subset M(x) + l\|x - x'\| B_F \quad \text{for all } x, x' \in X$$

and $M(\bar{x}) \neq \varnothing$, then for any $x \in X$

$$h(x, y^*) < \infty \Leftrightarrow h(\bar{x}, y^*) < \infty \Leftrightarrow h_r(\bar{x}, y^*) < \infty.$$

We set as in [10] (see also [2])

$$F_M^* = \{y^* \in F^*: h(\bar{x}, y^*) < \infty\}$$

and we easily see that F_M^* is a convex cone (the barrier cone of M see [2]).

PROPOSITION 5.2. *Assume that F is reflexive and endowed with a norm that is differentiable off zero, that $B_{F^*} \cap F_M^*$ is w^*-sequentially closed, and that M is locally Lipschitz around $\bar{x}$ (with $M(\bar{x}) \neq \varnothing$). Then, for $\Delta_r := \Delta_{M_r}$,*

$$(-x^*, y^*) \in \partial \Delta_r(\bar{x}, \bar{y}) \Rightarrow x^* \in \partial^\Box h_r(\cdot, y^*)(\bar{x}).$$

Proof. The set $S := B_{F^*} \cap F_M^*$ is w^*-sequentially compact and by (5.1)

$$\Delta_r(x, y) = \sup_{u^* \in S} (\langle u^*, y\rangle - h_r(x, u^*)).$$

Now, on the one hand, by Proposition 1.5 we have for any $u^* \in B_{F^*}$

$$\begin{aligned} h_r(\bar{x}, u^*) &= \sup \{\langle u^*, y'\rangle: y' \in M_r(\bar{x})\} \\ &= \sup \{\langle u^*, y'\rangle - d(y', M_r(\bar{x})): y' \in F\} \\ &= \sup \{\langle u^*, y'\rangle - \Delta_r(\bar{x}, y'): y' \in F\}. \end{aligned}$$

On the other hand, the function $\langle u^*, \cdot\rangle - \Delta_r(\bar{x}, \cdot)$ is concave (because $M(\bar{x})$ is convex) and $y^* \in \partial_2 \Delta_r(\bar{x}, \bar{y}) = \partial(d_{M_r(\bar{x})})(\bar{y})$ (see Proposition 5.3). So we obtain

$$0 \in \partial(\langle y^*, \cdot\rangle - \Delta_r(\bar{x}, \cdot))(\bar{y})$$

and

$$h_r(\bar{x}, y^*) = \langle y^*, \bar{y}\rangle - \Delta_r(\bar{x}, \bar{y}).$$

Therefore we have

$$h_r(\bar{x}, y^*) < \infty \quad \text{and} \quad y^* \in S(\bar{x}, \bar{y}),$$

which implies by the strict convexity of $h_r(\bar{x}, \cdot)$

$$S(\bar{x}, \bar{y}) = \{y^*\}.$$

Applying Proposition 5.1 and setting $q_{u^*}(x, y) = \langle u^*, y\rangle - h_r(x, u^*)$, we get

$$(-x^*, y^*) \in \partial^{\square} q_{y^*}(\bar{x}, \bar{y}) = (0, y^*) - \partial^{\square} h_r(\cdot, y^*)(\bar{x}) \times \{0\}$$

and hence

$$x^* \in \partial^{\square} h_r(\cdot, y^*)(\bar{x}). \qquad \square$$

Remark. If M is given by $M(x) = N(x) + Q$ where N is a multifunction which takes closed, convex, and bounded values and which is Lipschitz around $\bar{x}$ with $N(\bar{x}) \neq \emptyset$ and where Q is a closed convex cone in F, then it is not difficult to see that

$$F_M^* = Q^0 := \{y^* \in F^* : \langle y^*, y\rangle \leqq 0 \; \forall y \in Q\}.$$

Thus F_M^* is w^*-closed, which ensures that $B_{F^*} \cap F_M^*$ is w^*-sequentially compact whenever F is reflexive. $\square$

The following result, which is used in the proof of the proposition above, will also be needed in the sequel.

PROPOSITION 5.3. *Let $k \geqq 0$, and let $P: G \rightrightarrows H$ be a multifunction from a normed vector space G into the convex subsets of a normed space H. Assume that P is pseudo-Lipschitz at $(\bar{x}, \bar{y}) \in GrP$. Then there exists a neighbourhood $X \times Y$ of $(\bar{x}, \bar{y})$ such that for each $(x, y) \in X \times Y$ and $(x^*, y^*) \in t\, \partial\Delta_p(x, y)$ with $t > 0$, we have*

$$y^* \in t\, \partial d_{P(x)}(y).$$

Proof. Choose an open neighbourhood $X \times Y$ of $(\bar{x}, \bar{y})$ in $G \times H$ and a real number $\alpha > 0$ such that Δ_P is α-Lipschitz over $X \times Y$. As $\Delta_P(x, \cdot)$ is convex we obtain, by Proposition 4.4 of [9],

$$y^* \in \partial_2(t\Delta_P)(x, y) = t\partial d_{P(x)}(y). \qquad \square$$

The following proposition is an adaptation of the techniques introduced by Clarke [7] in order to give necessary conditions for optimal control problems in terms of the Hamiltonian (see also Clarke [6]).

PROPOSITION 5.4. *Let $(\bar{x}, \bar{u})$ be a local minimum point for the problem*

$$\min\{f(x, u) : u \in M(x), x \in A\}.$$

If f and the multifunction M are k-Lipschitz around $\bar{x}$, then for ε small enough and $0 < r < k^{-1}\varepsilon^2$ there exist $x_\varepsilon \in A$ and $u_\varepsilon \in M_r(x_\varepsilon)$ with

$$\|x_\varepsilon - \bar{x}\| \leqq \varepsilon, \quad \|u - \bar{u}\| \leqq \varepsilon, \quad f(\bar{x}, \bar{u}) \leqq f(x_\varepsilon, u_\varepsilon) \leqq f(\bar{x}, \bar{u}) + \varepsilon^2$$

such that $(x_\varepsilon, u_\varepsilon)$ is a local minimum of the function

$$(x, u) \to f(x, u) + \varepsilon d((x, u), (x_\varepsilon, u_\varepsilon)) + K d_A(x) + K\Delta_r(x, u),$$

where K is a constant independent of (ε, x, u).

Proof. Suppose that $(\bar{x} + sB) \times (\bar{u} + 2sB)$ is a neighbourhood over which $(\bar{x}, \bar{u})$ is a minimum for the above problem and over which f and M are k-Lipschitz. Consider

$\varepsilon < \min(s, (ks)^{1/2})$. Then for $x \in A \cap (\bar{x} + sB)$ and $u \in M_r(x) \cap (\bar{u} + sB)$ there exists $b \in B$ with $u - rb \in M(x) \cap (\bar{u} + 2sB)$ and hence

$$f(x, u) - f(\bar{x}, \bar{u}) - \varepsilon^2 \geqq f(x, u - rb) - f(\bar{x}, \bar{u}) + \varepsilon^2 - kr\|b\| \geqq \varepsilon^2 - kr > 0.$$

Therefore by Ekeland's variational principle (see [2]) there exists $x_\varepsilon \in A \cap (\bar{x} + sB)$ and $u_\varepsilon \in M_r(x_\varepsilon) \cap (\bar{u} + sB)$ such that

$$\|x_\varepsilon - \bar{x}\| \leqq \varepsilon, \quad \|u - \bar{u}\| \leqq \varepsilon, \quad f(\bar{x}, \bar{u}) \leqq f(x_\varepsilon, u_\varepsilon) \leqq f(\bar{x}, \bar{u}) + \varepsilon^2$$

and $(x_\varepsilon, u_\varepsilon)$ is a minimum point of the function

$$(x, u) \to f(x, u) - f(\bar{x}, \bar{u}) + \varepsilon d((x, u), (x_\varepsilon, u_\varepsilon))$$

over the set $\{(x, u) \colon u \in M_r(x) \cap (\bar{u} + sB), x \in A \cap (\bar{x} + sB)\}$. So it suffices to apply Proposition 1.5 to get the conclusion of the proposition. □

We can now give estimates of the subdifferential of the optimal value function in terms of the support function of M.

PROPOSITION 5.5. *Let $f \colon E \times F \to \mathbb{R}$ be locally Lipschitz, let $M \colon E \rightrightarrows F$ be a locally Lipschitz multifunction with closed convex values, and let A be a closed subset of E. Let*

$$m(u) = \inf_x \{f(x, u) \colon u \in M(x), x \in A\}.$$

Assume E admits an equivalent Fréchet differentiable norm (off zero), condition (K) is satisfied at $\bar{u}$, and $B_F^* \cap F_M^*$ is w^*-sequentially compact. Then for each $u^* \in \delta m(\bar{u})$ (respectively, $\delta^\infty m(\bar{u})$) there exists $\bar{x} \in S(\bar{u})$, $\lambda > 0$, and $(x^*, y^*) \in E^* \times F^*$ such that

$$(x^*, u^* - y^*) \in \delta(f + \lambda d_{A \times F})(\bar{x}, \bar{u}), \quad -x^* \in \partial^\square h(\cdot, y^*)(\bar{x}), \quad y^* \in \lambda \delta d_{M(\bar{x})}(\bar{u})$$

(respectively, $0 \in \partial^\square h(\cdot, u^*)(\bar{x}) + \gamma \delta d_A(\bar{x})$ and $u^* \in \lambda \partial \delta_{M(\bar{x})}(\bar{u})$).

Proof. Let $u^* \in \delta m(\bar{u})$. There exists $\varepsilon_n \downarrow 0$, $u_n \to \bar{u}$ with $m(u_n) \to m(\bar{u})$, $u_n^* \xrightarrow{w} \bar{u}$ such that the function

$$u \to m(u) - \langle u_n^*, u - u_n \rangle + \varepsilon_n \|u - u_n\|$$

attains a local minimum at u_n. By assumption (K) we may suppose that there exists $x_n \in S(u_n)$ with (x_n) converging to some $\bar{x} \in S(\bar{u})$. Then for all $x \in A$ and $u \in M(x)$ with u near u_n

$$\begin{aligned} -\langle u_n^*, u_n \rangle + f(x_n, u_n) &\leqq \varepsilon_n \|u - u_n\| - \langle u_n^*, u \rangle + m(u) \\ &\leqq \varepsilon_n \|u - u_n\| - \langle u_n^*, u \rangle + f(x, u). \end{aligned}$$

Therefore by Proposition 5.4 there exist a constant $\lambda > 0$ and $(x_n', u_n') \in (x_n, u_n) + \varepsilon_n B$, which is a local minimum of the function

$$(x, u) \to \varepsilon_n \|u - u_n\| - \langle u_n^*, u \rangle + f(x, u) + \varepsilon_n \|u - u_n'\| + \varepsilon_n \|x - x_n'\| + \lambda d_A(x) + \lambda \Delta_{r_n}(x, u),$$

and hence

$$(0, 0) \in 2\varepsilon_n B - (0, u_n^*) + \delta(f + \lambda d_{A \times F})(x_n', u_n') + \lambda \partial \Delta_{r_n}(x_n', u_n'),$$

which implies there exist $(z_n^*, v_n^*) \in \partial \Delta_{r_n}(x_n', u_n')$ satisfying

$$(-\lambda z_n^*, u_n^* - \lambda v_n^*) \in 2\varepsilon_n B + \delta(f + \lambda d_{A \times F})(x_n', u_n').$$

Extracting a subsequence if necessary we may suppose that $-\lambda z_n^* \xrightarrow{w} x^*$ and $\lambda v_n^* \xrightarrow{w} y^*$, which ensures $(x^*, u^* - y^*) \in \delta(f + \lambda d_{A \times F})(\bar{x}, \bar{u})$ and $x^* \in \partial^\square h(\cdot, y^*)(\bar{x})$, since by Proposition 5.2, we have $-\lambda z_n^* \in \partial^\square h_{r_n}(\cdot, \lambda v_n^*)(x_n')$.

Moreover, by Proposition 5.3 we have $v_n^* \in \partial d_{M_{r_n(x'_n)}}(u'_n)$ and by l-Lipschitz continuity of M around $\bar{x}$ we have

$$M_{r_n}(x'_n) \subset M(\bar{x}) + r_n B + l\|x'_n - \bar{x}\| B \quad \text{and} \quad M(\bar{x}) \subset M_{r_n}(\bar{x}) \subset N_{r_n}(x'_n) + l\|x'_n - \bar{x}\| B,$$

and hence for every $u \in F$

$$d_{M_{r_n}}(x'_n)(u) \geqq d_{M(\bar{x})}(u) - r - l\|x'_n - \bar{x}\| \quad \text{and} \quad d_{M(\bar{x})}(u) \geqq d_{M_{r_n}}(x'_n)(u) - l\|x'_n - \bar{x}\|.$$

Therefore for any $u \in F$ we have

$$\begin{aligned}\langle v_n^*, u - u'_n \rangle &\leqq d_{M_{r_n}}(x'_n)(u) - d_{M_{r_n}}(x'_n)(u'_n) \\ &\leqq d_{M(\bar{x})}(u) + l\|x'_n - \bar{x}\| - (d_{M(\bar{x})}(u'_n) - r_n - l\|x'_n - \bar{x}\|)\end{aligned}$$

and hence

$$\langle \lambda^{-1} y^*, u - \bar{u} \rangle \leqq d_{M(\bar{x})}(u) - d_{M(\bar{x})}(\bar{u}),$$

which ensures $y^* \in \lambda \partial d_{M(\bar{x})}(\bar{u}) = \lambda \delta_{M(\bar{x})}(\bar{u})$ and completes the proof for $u^* \in \delta m(\bar{u})$.

For $u^* \in \delta^\infty m(\bar{u})$ the proof is similar. □

Acknowledgment. The author is grateful to a referee for his comments.

REFERENCES

[1] J. B. AUBIN, *Lipschitz behavior of solutions to convex minimization problems*, Math. Oper. Res., 9 (1984), pp. 87-111.

[2] J. P. AUBIN AND I. EKELAND, *Applied Nonlinear Analysis*, Wiley-Interscience, New York, 1984.

[3] J. M. BORWEIN AND H. M. STROJWAS, *Proximal analysis and boundaries of closed sets in Banach space, part* I: *theory*, Canad. Math. J., 38 (1986), pp. 431-452.

[4] ———, *Proximal analysis and boundaries of closed sets in Banach space, part* II: *applications*, Canad. Math. J., 39 (1987), pp. 428-472.

[5] F. H. CLARKE, *Optimal solutions to differential inclusions*, J. Optim. Theory Appl., 19 (1976), pp. 469-479.

[6] ———, *Necessary conditions for a general control problem*, in Calculus of Variations and Control Theory, D. Russel, ed., Mathematics Research Center, Pub. 36, University of Wisconsin, Academic Press, New York, 1976, pp. 259-278.

[7] ———, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, 1983.

[8] 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.

[9] R. CORREA AND L. THIBAULT, *Subdifferential analysis of bivariate separately regular functions*, J. Math. Anal. Appl., 148 (1990), pp. 157-174.

[10] P. H. DIEN, *On the regularity condition for the extremal problem under locally Lipschitz inclusion constraints*, Appl. Math. Optim., 13 (1985), pp. 151-161.

[11] ———, *Locally Lipschitz set-valued mappings and general extremal problems*, Acta Math. Vietnam., 2 (1983), pp. 109-122.

[12] J. DIESTEL, *Geometry of Banach Spaces: selected topics*, Springer-Verlag, New York, 1975.

[13] S. DOLECKI, *Tangency and differentiation: marginal functions*, to appear.

[14] J. GAUVIN, *The generalized gradient of a marginal value function in mathematical programming*, Math. Oper. Res., 4 (1979), pp. 458-463.

[15] J. HAGLER AND H. SULLIVAN, *Smoothness and weak sequential compactness*, Proc. Amer. Math. Soc., 78 (1980), pp. 497-503.

[16] J. B. HIRIART-URRUTY, *Tangent cones, generalized gradients and mathematical programming in Banach spaces*, Math. Oper. Res., 4 (1979), pp. 79-97.

[17] A. D. IOFFE, *Approximate subdifferentials and applications* I: *The finite dimensional theory*, Trans. Amer. Math. Soc., 281 (1984), pp. 389-416.

[18] ———, *On subdifferentiability spaces*, Ann. New York Acad. Sci., 410 (1983), pp. 107-120.

[19] A. JOURANI AND L. THIBAULT, *Approximate subdifferential and metric regularity: the finite dimensional case*, Math. Programming, 47 (1990), pp. 203-218.

[20] A. Y. KRUGER, *Properties of generalized differentials*, Siberian Math. J., (1983), pp. 822-832.

[21] A. Y. KRUGER AND B. S. MORDUKHOVICH, *Extreme points and the euler equations in nondifferentiable optimization problems*, Dokl. Akad. Nauk BSSR, 24 (1980), pp. 684-687.

[22] L. McLinden, *An application of Ekeland's theorem to minimax problems*, Nonlinear Anal. Theory Methods Appl., 6 (1982), pp. 189-196.

[23] B. S. Mordukhovich, *Maximum principle in the optimal time control problem with non-smooth constraints*, J. Appl. Math. Mech., 40 (1976), pp. 960-969.

[24] ———, *Metric approximations and necessary optimality conditions for general classes of nonsmooth extremal problems*, Soviet Math. Dokl., 22 (1980), pp. 526-530.

[25] J. P. Penot, *On regularity conditions in mathematical programming*, Math. Programming, 19 (1982), pp. 167-193.

[26] S. M. Robinson, *Stability theory for systems of inequalities, part* II: *differentiable nonlinear systems*, SIAM J. Numer. Anal., 13 (1976), pp. 497-513.

[27] R. T. Rockafellar, *Directionally Lipschitzian functions and subdifferential calculus*, Proc. London Math. Soc. (3), 39 (1979), pp. 331-355.

[28] ———, *Proximal subgradients, marginal values, and augmented Lagrangians in nonconvex optimization*, Math. Oper. Res., 6 (1981), pp. 427-437.

[29] ———, *Lagrange multipliers and subderivatives of optimal value functions in nonlinear programming*, Math. Programming Stud., 17 (1982), pp. 28-66.

[30] ———, *Directional differentiability of the optimal value function in a nonlinear programming problem*, Math. Programming Stud., 21 (1984), pp. 213-226.

[31] ———, *Extensions of subgradient calculus with applications to optimization*, Nonlinear Anal. Theory Methods Appl., 9 (1985), pp. 665-698.

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

[33] L. Thibault, *On generalized differentials and subdifferentials of Lipschitz vector-valued functions*, Nonlinear Anal. Theory Methods Appl., 6 (1982), pp. 1037-1053.

[34] J. S. Treiman, *A new characterization of Clarke's tangent cone and its application to subgradient analysis and optimization*, Ph.D. thesis, University of Washington, Seattle, WA, 1983.

[35] ———, *Clarke's gradients and epsilon-subgradients in Banach spaces*, Trans. Amer. Math. Soc., 294 (1986), pp. 65-78.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 5, pp. 1037-1060, September 1991

ON THE CONVERGENCE OF A MATRIX SPLITTING ALGORITHM FOR THE SYMMETRIC MONOTONE LINEAR COMPLEMENTARITY PROBLEM*

ZHI-QUAN LUO†‡ AND PAUL TSENG†

Abstract. A matrix splitting algorithm for the linear complementarity problem is considered, where the matrix is symmetric positive semidefinite. It is shown that if the splitting is regular, then the iterates generated by the algorithm are well defined and converge to a solution. This result resolves in the affirmative a long standing question about the convergence of the point successive overrelaxation (SOR) method for solving this problem. This result is also extended to related iterative methods. As direct consequences, convergence of the methods of, respectively, Aganagic, Cottle et al., Mangasarian, Pang, and others, is obtained, without making any additional assumptions on the problem.

Key words. convergence, iterative methods, convex quadratic program, linear complementarity, matrix splitting, gradient projection, SOR

AMS(MOS) subject classifications. 49, 90

1. Introduction.

Let M be an $n \times n$ symmetric positive semidefinite matrix and let q be an element of $\Re^n$ (the n-dimensional Euclidean space). Consider the following convex quadratic program:

$$\text{(P)} \qquad \begin{array}{ll} \text{minimize} & f(x) = \tfrac{1}{2}\langle x, Mx\rangle + \langle q, x\rangle \\ \text{subject to} & x \geqq 0. \end{array}$$

In our notation, all vectors are column vectors, $\langle \cdot, \cdot \rangle$ denotes the usual Euclidean inner product, and, for any vector x, x_i denotes its ith coordinate. We note that all our results can be extended in a straightforward manner to problems with general box constraints of the form

$$l \leqq x \leqq c,$$

where l is any element of $[-\infty, \infty)^n$ and c is any element of $(-\infty, \infty]^n$ (e.g., unconstrained problems). However, in order to simplify the presentation, we will not treat the more general problems here.

The problem (P), commonly referred to as the *symmetric monotone linear complementarity* problem, has a number of important applications to, for example, linear/quadratic programming [BeT89], [Man77], [MaD87], [LiP87] and the solution of certain boundary value problems [CoG78], [CGS78], [DeT84].

We make the following standing assumption on (P):

Assumption A. f is bounded from below on the feasible set $X = [0, \infty)^n$.

Since f is convex quadratic and X is a polyhedral set, it follows from a standard result in quadratic programming (see, e.g., [Eav71], [FrW57]) that (P) has a finite optimal value and the set of optimal solutions for (P), denoted by X^*, is nonempty. However, because M is only positive semidefinite, X^* may be unbounded.

* Received by the editors August 8, 1989; accepted for publication (in revised form) July 27, 1990. This research was partially supported by United States Army Research Office contract DAAL03-86-K-0171 (Center for Intelligent Control Systems), by National Science Foundation grant NSF-ECS-8519058, and by a grant from the Science and Engineering Research Board of McMaster University.

† Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

‡ Present address, Department of Electrical and Computer Engineering, McMaster University, Room 225/CRL, Hamilton, Ontario, Canada L8S 4L7.

From the Kuhn-Tucker conditions for (P) it is easily seen that an x belongs to X^* if and only if the orthogonal projection of $x-\nabla f(x)$ onto the feasible set X is x itself, i.e.,

$$x=[x-(Mx+q)]^+, \tag{1.1}$$

where $[y]^+$ denotes the orthogonal projection of y onto X. Now, let us write M as

$$M=B+C, \tag{1.2}$$

for some $n\times n$ matrices B and C. In the terminology of numerical analysis [OrR70], such a pair (B, C) is called a *splitting* of M. If in addition $B-C$ is positive definite (not necessarily symmetric), then (B, C) is called a *regular* splitting of M (cf. [LiP87]).

Suppose that, instead of solving the nonlinear equation (1.1) directly, we fix a solution estimate $x\in X$ and solve the following approximation to (1.1)

$$y=[y-(By+Cx+q)]^+, \tag{1.3}$$

to obtain a solution y. We can then set x to y and repeat the procedure. We formalize this procedure with the following iterative scheme. Let (B, C) be a regular splitting of M. Define a corresponding point-to-point mapping $\mathcal{A}_B: X\mapsto X$ by (cf. (1.3))

$$\mathcal{A}_B(x)=\{y\in\Re^n \mid y=[y-(By+Cx+q)]^+\} \quad \forall x\in X. \tag{1.4}$$

We show in § 2 that $\mathcal{A}_B$ is well defined (see Lemma 2(a)). Note that an x is an optimal solution of (P) if and only if $x=\mathcal{A}_B(x)$. Consider the following algorithm for solving (P).

MATRIX SPLITTING ALGORITHM. Choose an $x^0\in X$. Generate a sequence of vectors $\{x^0, x^1, \cdots\}$ in X by the formula

$$x^{r+1}=\mathcal{A}_B(x^r), \qquad r=0,1,\cdots. \tag{1.5}$$

In order for the algorithm (1.4)-(1.5) to be practical, the splitting (B, C) should be chosen so that (1.3) is easily solvable. We will discuss such choices in § 5.

The first matrix splitting algorithm for solving the problem (P) is the cyclic coordinate descent method of Hildreth [Hil57]. This method is simple, uses little storage, can exploit problem sparsity, and is practical for solving large scale problems. The method of Hildreth was subsequently extended by Cryer [Cry71] to a (point) SOR method, which in turn was extended by Cottle, Golub, and Sacher [CGS78] and Cottle and Pang [CoP82] to block successive overrelaxation (SOR) methods. Cottle and Goheen [CoG78] further extended the Cottle-Golub-Sacher method to solve problems with box constraints. An extension of Cryer's method along a different direction was proposed by Mangasarian [Man77], which is also closely related to a gradient projection algorithm of Aganagic [Aga78]. (Applications of Mangasarian's method to solving strictly convex quadratic programs and linear programs are discussed in [Man84] and [MaD87]. Parallel implementation of the method is discussed in [MaD87].) Pang [Pan82] showed that the above methods (with the possible exception of the block SOR methods) can be viewed as special cases of the matrix splitting algorithm (1.4)-(1.5). Pang then proceeded to give an extensive analysis of this algorithm [Pan82], [Pan84], [Pan86]. Yet, despite their long history and practical advantages, convergence of these iterative methods remained largely unresolved. (A summary of the current knowledge is given in [LiP87, § 2-3]. See [BeT89, Chap. 3] and [Che84] for discussions on gradient projection algorithms.) In particular, none of the above methods has been shown to be convergent (in the sense that the iterates converge to an optimal solution) if the optimal solution is not unique. Convergence typically requires additional assumptions on the problem, all of which lead to the compactness of the solution set X^*, in which case the proof becomes rather routine (i.e., checking that each limit point is an optimal solution). In the absence of any such assumption, it was only known that the gradient of the iterates converge and that each limit point of the iterates, if it exists, is an optimal solution. The method of Cottle and Pang [CoP82] does generate a limit point, but this method includes, in addition to the standard block SOR iteration, a projection step to ensure that the iterates stay bounded and, moreover, it is applicable only to problems with a network structure. It is the aim of this paper to resolve this fundamental question of convergence by showing that the above methods are indeed convergent without making any additional assumptions on the problem. In fact, we prove a more general result that, if the splitting is regular, then the corresponding matrix splitting algorithm (1.4)–(1.5) is well defined and convergent, and the same conclusion holds for certain SOR extensions of the algorithm. (To the best of our knowledge, the only other matrix splitting algorithm that is known to be convergent in the same strong sense is one considered in Tseng [Tse89].[1]) Our proof is of some interest in itself as it uses certain (new) contraction properties of regular splitting and gives a detailed analysis of the trajectory of the iterates near the boundary of the feasible set X.

We remark that even for the simplest instance of the matrix splitting algorithm (1.4)–(1.5), such as the cyclic coordinate descent method, convergence is very difficult to establish when the cost function has unbounded level sets. The only other nontrivial problem having unbounded level sets in the cost function, and for which the cyclic coordinate descent method is known to be convergent in our strong sense, is a certain dual problem arising in nonlinear network optimization [BHT87].

This paper proceeds as follows. In § 2 we derive a number of properties of the solutions of (P) and of regular splitting. In § 3 we use these properties to prove that, when the splitting is regular, the iterates generated by the algorithm (1.4)–(1.5) converge to an optimal solution of (P). In § 4 we propose SOR extensions of this algorithm. In § 5 we apply the above results to a number of known methods.

In our notation, superscript T will denote matrix transpose and $\|\cdot\|$, $\|\cdot\|_\infty$ will denote, respectively, the L_2-norm and the L_∞-norm in some Euclidean space. If A is a square matrix, $\|A\|$ will denote the matrix norm of A induced by the vector norm $\|\cdot\|$, i.e., $\|A\|=\max_{\|x\|=1}\|Ax\|$. For any $k\times m$ matrix A, we will denote by A_i the ith row of A and, for any nonempty $I\subseteq\{1,\cdots,k\}$ and $J\subseteq\{1,\cdots,m\}$, by A_I the submatrix of A obtained by removing all rows i of A such that $i\notin I$, and by A_{IJ} the submatrix of A_I obtained by removing all columns j of A_I such that $j\notin J$. We will also denote by Span (A) the space spanned by the columns of A. Analogously, for any k-vector x and any nonempty subset $J\subseteq\{1,\cdots,k\}$, we denote by x_J the vector with components x_i, $i\in J$. For any finite set J, we denote by Card (J) the cardinality of J. Finally, for any $J\subseteq\{1,\cdots,n\}$, we denote by $\tilde{J}$ the complement of J with respect to $\{1,\cdots,n\}$.

2. Characterization of optimal solutions and regular splittings. In this section we derive various properties of the elements of X^* and the mapping $\mathcal{A}_B$ given by regular splittings (B, C) of M. These properties will be used in the following section to prove convergence of the algorithm (1.4)–(1.5).

The first result, which is well known (see [AdG75] or [Man88]), states that ∇f is invariant over the solution set X^*.

[1] When the splitting is regular and *symmetric*, a simpler proof of convergence for the matrix splitting algorithm with inexact subproblem solution was recently found by Mangasarian (see [Man90]).

Lemma 1. *There exists a $d^* \in \Re^n$ such that $Mx^* + q = d^*$ for all $x^* \in X^*$.*

The next result shows that, if (B, C) is a regular splitting of M, then $\mathcal{A}_B$ is a well-defined point-to-point mapping and possesses a certain descent property.

Lemma 2. *Let (B, C) be a regular splitting of M. Then the following hold:*

(a) *$\mathcal{A}_B : X \mapsto X$ is a well-defined point-to-point mapping.*

(b) *For any $x \in X$,*

$$f(y) - f(x) \leqq \langle y - x, (C - B)(y - x)\rangle / 2,$$

where $y = \mathcal{A}_B(x)$.

Proof. We first prove part (a). Since $B - C$ is positive definite, it follows from $2B = M + (B - C)$ (cf. $M = B + C$) and the positive semidefinite property of M that B is positive definite. Hence, by a well-known result on variational inequality [BeT89, p. 271], [KiS80, § 2], we have that, for any $x \in X$, the nonlinear equation

$$y = [y - (By + Cx + q)]^+,$$

has a unique solution y. This proves part (a).

Now we prove part (b). It can be seen by using $M = B + C$ that, for any x and y in $\Re^n$,

$$f(y) - f(x) = \langle y - x, By + Cx + q\rangle + \langle y - x, (C - B)(y - x)\rangle / 2.$$

On the other hand, we see from (1.4) that $y = \mathcal{A}_B(x)$ if and only if $y \in X$ and

$$B_i y + C_i x + q_i \geqq 0, \quad B_i y + C_i x + q_i > 0 \Rightarrow y_i = 0 \quad \forall i.$$

Hence if in addition $x \in X$ (so that $x \geqq 0$), then $\langle y - x, By + Cx + q\rangle \leqq 0$. □

(The results of Lemma 2 are quite well known (e.g., [LiP87]). The proof of part (b) is based on one given in Lemma 4.1 of [Pan84].)

It can be seen that if the nonnegativity constraints $x \in X$ are removed, then, for any splitting (B, C) of M, $y = \mathcal{A}_B(x)$ if and only if $By + Cx + q = 0$, or equivalently (assuming that B is invertible)

$$y = -B^{-1}(Cx + q) = (I - B^{-1}M)x - B^{-1}q.$$

We next study an important contractive property of the iteration matrix $I - B^{-1}M$ when the splitting is regular.

Definition 1. Let Q be an $n \times n$ real symmetric matrix and let Null (Q) be the null space of Q. (Clearly, $\Re^n$ is the direct sum of Null (Q) and Span (Q).) A matrix T of size $n \times n$ is said to be *convergent* for Q if T, when viewed as a linear transformation from $\Re^n$ to $\Re^n$, is equal to identity in Null (Q) *and* is contractive in Span (Q). In other words, T satisfies

(1) $Tz = z$ for all $z \in$ Null (Q);

(2) For any $z \in$ Span (Q), the sequence $\{\|T^k z\|\}$ converges to zero geometrically, as $k \to \infty$ (with a convergence ratio that depends on T only).

The following important result is due to Keller [Kel65, Thm. 2] (see also [BeP79, p. 201]).

Proposition 1. *Let Q be a real symmetric matrix and let N be a nonsingular matrix for which the matrix*

$$N + N^T - Q \tag{2.1}$$

is positive definite. Then $T = I - N^{-1}Q$ is convergent for Q if and only if Q is positive semidefinite.

As a direct consequence of Keller's result, we have the following contractive properties of the iteration matrix $I - B^{-1}M$, for any regular splitting (B, C) of M.

LEMMA 3. *Let Q be an $m \times m$ symmetric positive semidefinite matrix and let (N, H) be a regular splitting of Q. Then N is positive definite and the following hold:*

(a) *There exist $\rho \in (0, 1)$ and $\tau > 0$ such that*

$$\left\| \prod_{h=1}^{k} (I - \theta^h Q N^{-1}) z \right\| \leqq \tau(1 - \underline{\theta}(1-\rho))^k \|z\| \quad \forall k \geqq 1, \forall z \in \operatorname{Span}(Q),$$

for any $\underline{\theta} \in (0, 1]$ and any sequence of scalars $\{\theta^1, \theta^2, \cdots\}$ in the interval $[\underline{\theta}, 1]$.

(b) *There exists $\Delta \geqq 1$ such that*

$$\left\| \prod_{h=1}^{k} (I - \theta^h N^{-1} Q) z \right\| \leqq \Delta \|z\| \quad \forall k \geqq 1, \quad \forall z \in \mathfrak{R}^m,$$

for any $\underline{\theta} \in (0, 1]$ and any sequence of scalars $\{\theta^1, \theta^2, \cdots\}$ in the interval $[\underline{\theta}, 1]$.

Proof. From $Q = N + H$ we have $N - H = 2N - Q$, so the symmetric part of $N - H$ is $N + N^T - Q$. Since $N - H$ is positive definite so its symmetric part is also positive definite, then Proposition 1 yields that $I - N^{-1}Q$ is convergent for Q. This implies

$$(2.2) \qquad (I - N^{-1}Q) z \in \operatorname{Span}(Q) \quad \forall z \in \operatorname{Span}(Q),$$

and that there exists a $\rho \in (0, 1)$ depending on N and Q only such that

$$(2.3) \qquad \|(I - N^{-1}Q) z\| \leqq \rho \|z\| \quad \forall z \in \operatorname{Span}(Q).$$

(a) Consider any $z \in \operatorname{Span}(Q)$, any integer $k \geqq 1$, any $\underline{\theta} \in (0, 1]$, and any sequence of scalars $\{\theta^1, \theta^2, \cdots\}$ in the interval $[\underline{\theta}, 1]$. Since Q is symmetric positive semidefinite, then there exists a w satisfying $z = Qw$ and

$$(2.4) \qquad \|w\| \leqq \frac{1}{\sigma} \|z\|,$$

where σ denotes the smallest nonzero eigenvalue of Q. Let w^0 denote the orthogonal projection of w onto $\operatorname{Span}(Q)$ and let $w^1, \cdots, w^k$ be given by the formula

$$(2.5) \qquad w^h = (I - \theta^h N^{-1} Q) w^{h-1}, \qquad h = 1, \cdots, k.$$

Then, $z = Qw^0$, $w^0 \in \operatorname{Span}(Q)$, and it is easily seen by using induction on h (together with (2.2), (2.5)) that

$$w^h \in \operatorname{Span}(Q) \quad \forall h = 0, 1, \cdots, k.$$

Then, for each $h \in \{1, \cdots, k\}$, since $w^{h-1} \in \operatorname{Span}(Q)$, we have from (2.3) and (2.5) that

$$\begin{aligned}
\|w^h\| &= \|(I - \theta^h N^{-1} Q) w^{h-1}\| \\
&= \|(1 - \theta^h) w^{h-1} + \theta^h (I - N^{-1} Q) w^{h-1}\| \\
(2.6) \qquad &\leqq (1 - \theta^h) \|w^{h-1}\| + \theta^h \|(I - N^{-1} Q) w^{h-1}\| \\
&\leqq (1 - \theta^h) \|w^{h-1}\| + \theta^h \rho \|w^{h-1}\| \\
&\leqq (1 - \underline{\theta}(1 - \rho)) \|w^{h-1}\|,
\end{aligned}$$

where the last inequality follows from $\theta^h \geqq \underline{\theta}$, and hence

$$\|w^k\| \leqq (1 - \underline{\theta}(1-\rho))^k \|w^0\| \leqq (1 - \underline{\theta}(1-\rho))^k \|w\|.$$

This combined with the observation $\prod_{h=1}^{k}(I-\theta^h QN^{-1})z = Qw^k$ (cf. (2.5) and $z=Qw^0$) and (2.4) then yields

$$\begin{aligned}\left\|\prod_{h=1}^{k}(I-\theta^h QN^{-1})z\right\| &= \|Qw^k\| \\ &\leqq \|Q\|(1-\underline{\theta}(1-\rho))^k\|w\| \\ &\leqq \frac{\|Q\|}{\sigma}(1-\underline{\theta}(1-\rho))^k\|z\|.\end{aligned} \tag{2.7}$$

By setting $\tau=\|Q\|/\sigma$, we complete the proof of part (a).

(b) Consider any $z\in\Re^n$, any integer $k\geqq 1$, any $\underline{\theta}\in(0,1]$, and any sequence of scalars $\{\theta^1,\theta^2,\cdots\}$ in the interval $[\underline{\theta},1]$. Let us decompose z into the sum of z' and z'', for some $z'\in\operatorname{Span}(Q)$ and some $z''\in\operatorname{Null}(Q)$. Since $I-N^{-1}Q$ is convergent for Q we have $(I-N^{-1}Q)z''=z''$, so that

$$(I-\theta^h N^{-1}Q)z''=(1-\theta^h)z''+\theta^h(I-N^{-1}Q)z''=z'',\qquad h=1,\cdots,k.$$

Hence

$$\prod_{h=1}^{k}(I-\theta^h N^{-1}Q)z''=z''. \tag{2.8}$$

Let $z^0=z'$ and let $z^1,\cdots,z^k$ be given by the formula

$$z^h=(I-\theta^h N^{-1}Q)z^{h-1},\qquad h=1,\cdots,k.$$

Then, $z^0\in\operatorname{Span}(Q)$ and it is easily seen by using induction on h (together with (2.2)) that $z^h\in\operatorname{Span}(Q)$ for all h, so (2.3) yields

$$\begin{aligned}\|z^h\| &= \|(I-\theta^h N^{-1}Q)z^{h-1}\| \\ &\leqq (1-\theta^h)\|z^{h-1}\|+\theta^h\|(I-N^{-1}Q)z^{h-1}\| \\ &\leqq (1-\underline{\theta}(1-\rho))\|z^{h-1}\|,\qquad h=1,\cdots,k.\end{aligned} \tag{2.9}$$

Hence

$$\left\|\prod_{h=1}^{k}(I-\theta^h N^{-1}Q)z'\right\| = \|z^k\|\leqq(1-\underline{\theta}(1-\rho))^k\|z^0\|=(1-\underline{\theta}(1-\rho))^k\|z'\|.$$

By combining the above relation with (2.8) and using the obvious facts $\|z'\|\leqq\|z\|$, $\|z''\|\leqq\|z\|$, we obtain

$$\begin{aligned}\left\|\prod_{h=1}^{k}(I-\theta^h N^{-1}Q)z\right\| &\leqq \left\|\prod_{h=1}^{k}(I-\theta^h N^{-1}Q)z'\right\|+\|z''\| \\ &\leqq (1-\underline{\theta}(1-\rho))^k\|z\|+\|z\|\leqq 2\|z\|.\end{aligned}$$

□

Remark 1. Since $I-N^{-1}Q=N^{-1}(I-QN^{-1})N$, the two matrices $I-N^{-1}Q$ and $I-QN^{-1}$ are similar and therefore have identical eigenvalues. Hence part (b) of Lemma 3 implies that the eigenvalues of $I-QN^{-1}$ are also either inside or on the unit circle.

Remark 2. The relaxation parameters $\theta^1,\theta^2,\cdots$ are not needed for establishing our main result (Theorem 1), but they will be used in § 4 when we introduce under/over-relaxation to the mapping $\mathcal{A}_B$.

Remark 3. It can be seen by using Lemma 3(a) that, for any $b \in \operatorname{Span}(Q)$ and any $y^0 \in \Re^m$, the sequence of points $\{y^r\}$ generated according to

$$y^{r+1} = (I - N^{-1}Q)y^r - N^{-1}b, \qquad r = 0, 1, \cdots$$

converges geometrically.

Consider the coordinate descent method for solving the unconstrained version of (P) (i.e., find an x satisfying $Mx + q = 0$)

$$y^{r+1} = (I - (E+L)^{-1}M)y^r - (E+L)^{-1}q,$$

where E and L denote, respectively, the diagonal and the strictly lower triangular part of M. We assume that $q \in \operatorname{Span}(M)$ so the problem has a solution and that E has positive diagonal entries so the above iterations are well defined. (Since M is symmetric positive semidefinite, then a diagonal entry of M is zero if and only if the entire row (column) of M containing that entry is zero, so the second assumption is really not restrictive.) As an immediate consequence of Remark 3, we have that the iterates $\{y^r\}$ converge at a geometric rate. This is because the splitting $(E + L, L^T)$ is easily verified to be regular, so since $q \in \operatorname{Span}(M)$, Remark 3 then implies the geometric rate of convergence of $\{y^r\}$ to some x satisfying $Mx + q = 0$.

Lemma 3 in turn implies the following facts.

LEMMA 4. *Let (B, C) be a regular splitting of M. Then the following hold*:

(a) *For any nonempty $J \subseteq \{1, \cdots, n\}$, there exist $\rho_J \in (0, 1)$ and $\tau_J > 0$ such that*

$$\|(I - M_{JJ}(B_{JJ})^{-1})^k z\| \leqq \tau_J(\rho_J)^k\|z\| \quad \forall k \geqq 1, \quad \forall z \in \operatorname{Span}(M_{JJ}).$$

(b) *There exists a $\Delta \geqq 1$ such that, for any nonempty $J \subseteq \{1, \cdots, n\}$,*

$$\|(I - (B_{JJ})^{-1}M_{JJ})^k z\| \leqq \Delta\|z\| \quad \forall k \geqq 1, \quad \forall z.$$

Proof. Since $B - C$ is positive definite, $B_{JJ} - C_{JJ}$ is positive definite. Parts (a) and (b) then follow immediately from, respectively, parts (a) and (b) of Lemma 3. □

Let

$$I^* = \{i \in \{1, \cdots, n\} \mid d_i^* = 0\}.$$

Then, for each $x^* \in X^*$, we have $M_{I^*}x^* + q_{I^*} = 0$ (cf. Lemma 1), so that

$$q_{I^*} \in \operatorname{Span}(M_{I^*}). \tag{2.10}$$

Moreover, from (1.1) and Lemma 1 we have that $x^* = [x^* - d^*]^+$ for all $x^* \in X^*$. Since $[\cdot]^+$ is the orthogonal projection onto the nonnegative orthant, this shows that, for all $i \notin I^*$,

$$d_i^* > 0 \quad \text{and} \quad x_i^* = 0 \quad \forall x^* \in X^*. \tag{2.11}$$

In the remainder of this paper, we assume that $I^* \neq \emptyset$ for otherwise it is well known that the matrix splitting algorithm terminates finitely. The submatrix of M indexed by I^* has a number of interesting properties, which we show below.

LEMMA 5. *For any $J \subseteq I^*$,* $\operatorname{Span}(M_{J\tilde{J}}) \subseteq \operatorname{Span}(M_{JJ})$ *and* $q_J \in \operatorname{Span}(M_{JJ})$. (*Recall that $\tilde{J}$ denotes the complement of J with respect to* $\{1, \cdots, n\}$.)

Proof. The proof is by contradiction. Suppose that for some $i \notin J$, $M_{Ji} \notin \operatorname{Span}(M_{JJ})$. It then follows that there exists some vector $u \in \Re^{\operatorname{Card}(J)}$ such that

$$\langle u, M_{Ji}\rangle < 0, \qquad M_{JJ}u = 0. \tag{2.12}$$

Let x be the n-dimensional vector given by $x_J = u$, $x_i = 1$ and $x_j = 0$ for all $j \notin J$ with $j \neq i$. Then

$$\langle x, Mx\rangle = \langle u, M_{JJ}u\rangle + 2\langle u, M_{Ji}\rangle,$$

and it follows from (2.12) that

$$\langle x, Mx\rangle < 0,$$

a contradiction of the positive semidefinite property of M. Thus, Span $(M_{J\bar{J}}) \subseteq$ Span (M_{JJ}). This, together with the fact $q_J \in$ Span (M_J) (cf. (2.10) and $J \subseteq I^*$), implies $q_J \in$ Span (M_{JJ}). □

3. A general convergence theorem. Let $\{x^r\}$ be a sequence of iterates generated by the algorithm (1.4)-(1.5), i.e.,

$$x^{r+1} = \mathcal{A}_B(x^r), \qquad r = 0, 1, \cdots,$$

where (B, C) is some regular splitting of M. By Lemma 2(a), $\{x^r\}$ is well defined. We will show that $\{x^r\}$ converges to an element of X^*.

To motivate our proof, note from Lemma 2(b) that, for all r,

$$\begin{aligned} f(x^{r+1}) &\leqq f(x^r) - \langle x^{r+1} - x^r, (B - C)(x^{r+1} - x^r)\rangle/2 \\ &\leqq f(x^r) - \gamma\|x^{r+1} - x^r\|^2/2, \end{aligned} \tag{3.1}$$

where $\gamma > 0$ denotes the smallest eigenvalue of the symmetric part of $(B - C)$. Upon summing this inequality over all r and using the fact that $f(x^r)$ is bounded from below for all r (cf. Assumption A) and the fact that $\gamma > 0$, we obtain

$$\sum_{r=0}^{\infty} \|x^{r+1} - x^r\|^2 < \infty. \tag{3.2}$$

Hence $x^{r+1} - x^r \to 0$, which together with

$$x^{r+1} = [x^{r+1} - (Bx^{r+1} + Cx^r + q)]^+ \tag{3.3}$$

(cf. $x^{r+1} = \mathcal{A}_B(x^r)$), the Lipschitz continuity of $[\cdot]^+$, and the fact $B + C = M$, establishes the following lemma.

LEMMA 6. (a) $x^{r+1} - x^r \to 0$.

(b) $x^r - [x^r - Mx^r - q]^+ \to 0$.

Hence any limit point x^∞ of $\{x^r\}$ satisfies $x^\infty = [x^\infty - Mx^\infty - q]^+$ and is therefore in X^*. This result is quite well known (e.g., [Pan86], [LiP87]) and, as we just saw, is relatively easy to prove. The difficulty lies in showing that $\{x^r\}$ indeed has a limit point. This is a highly nontrivial task to which the remainder of this section will be devoted.

Remark 4. Equation (3.2) gives an estimate of the rate at which $x^{r+1} - x^r \to 0$, but technically speaking, it is not enough for us to claim the convergence of $\{x^r\}$ since it does not prevent $x^{r+1} - x^r$ to decrease like $1/r$, in which case $\|x^r\| \to \infty$. Intuitively, it seems unlikely that such a sequence of iterates can be generated by the matrix splitting algorithm, but to show this rigorously is very difficult, as indicated by the complexity of the proof given below.

For each $x \in \mathfrak{R}^n$, let $\phi(x)$ denote the distance from x to X^*, i.e.,

$$\phi(x) = \min_{x^* \in X^*} \|x - x^*\|.$$

The next lemma, which shows that $\{Mx^r\}$ converges and that $\{x^r\}$ comes arbitrarily close to X^*, follows as a direct consequence of Lemma 6.

LEMMA 7. (a) $Mx^r + q \to d^*$.

(b) $\phi(x^r) \to 0$.

Proof. Part (a) is shown in Theorem 3.1 of [Pan86]. Part (b) is a direct consequence of part (a) and the upper-Lipschitzian property of the solution set of a monotone linear complementarity problem (see [Rob81]). □

Now let us map out the directions for the most intricate part of our proof. By Lemma 7(b), we know that $\{x^r\}$ comes arbitrarily close to X^*, but we do not know if it is bounded. Now, it is easily seen from Lemmas 6(b) and 7(a) that those coordinates x_i^r, $i \notin I^*$, stay fixed to the boundary point 0 for all r sufficiently large; so we need to consider only those coordinates of x^r indexed by I^*. If these coordinates all stay strictly away from zero (which, for example, holds if every element x^* of X^* is *nondegenerate* in the sense that $x_i^* > 0$ for all $i \in I^*$), then the problem effectively becomes unconstrained and it immediately follows from Keller's result (see Remark 3) that all coordinates of x^r converge at a geometric rate. Hence the difficulty lies with those coordinates x_i^r $(i \in I^*)$ that bounce around the boundary point zero, possibly causing one of the remaining coordinates to sail off to infinity. To resolve this difficulty, we will show that these coordinates perturb the movement of the remaining coordinates only (additively) by their own maximum deviation from the boundary. This fact, shown in Lemma 8 below, is based on the contraction property of the algorithmic mapping for the unconstrained case (cf. Lemma 4) and Lemma 5. Then those coordinates of x^r that start out far from the boundary will stay far from the boundary (cf. geometric convergence for the unconstrained case), unless one of the remaining coordinates also moves far from the boundary, so that, eventually, each coordinate of x^r either stays close to the boundary or stays far from the boundary. Those coordinates that stay close to the boundary are clearly bounded; those coordinates that stay far from the boundary are also bounded because perturbation by the other coordinates is bounded and, within themselves, the convergence is geometric (since they are effectively unconstrained). We now proceed with the actual proof.

Let

$$\beta = 1 + \max_{J \subseteq I^*} \sqrt{\operatorname{Card}(\tilde{J})} \left\{ \left(\frac{\tau_J \|(B_{JJ})^{-1}\| \, \|M_{JJ}\|}{1-\rho_J} + \Delta + 1 \right) \|(B_{JJ})^{-1} B_{J\tilde{J}}\| + \frac{\tau_J \|(B_{JJ})^{-1}\| \, \|M_{J\tilde{J}}\|}{1-\rho_J} \right\}.$$

The following lemma, based on Lemmas 1, 4, and 5, shows that those coordinates of x^r that stay away from the boundary of X are influenced by the remaining coordinates only through the distance, scaled by β, of these remaining coordinates from the boundary of X. This result allows us to separate the effect of these two sets of coordinates on each other.

LEMMA 8 (coordinate separation). *Consider any $J \subseteq I^*$. If for some two integers $s \geqq t \geqq 0$ we have $x_i^r > 0$ for all $i \in J$ and all $r = t+1, t+2, \cdots, s$, then, for any $x^* \in X^*$,*

$$\|x_J^s - x_J^*\| \leqq \Delta \|x_J^t - x_J^*\| + \beta \max_{r \in \{t, \cdots, s\}} \|x_{\tilde{J}}^r - x_{\tilde{J}}^*\|_\infty .$$

Proof. The claim clearly holds if $s = t$ (since $\Delta \geqq 1$). Suppose that $s > t$. Since $x_i^r > 0$ for all $i \in J$ and all $r = t+1, \cdots, s$, it follows from the fact $x^{r+1} = [x^{r+1} - (Bx^{r+1} + Cx^r + q)]^+$ for all r (cf. (3.3)) that

$$B_J x^{r+1} + C_J x^r + q_J = 0, \qquad r = t, \cdots, s-1,$$

or equivalently (using $M_J = B_J + C_J$),

$$B_J(x^{r+1} - x^r) + M_J x^r + q_J = 0, \qquad r = t, \cdots, s-1.$$

Since $J \subseteq I^*$, we also have (using Lemma 1 and the definition of I^*)

$$M_J x^* + q_J = 0.$$

Combining the above two equalities and multiplying by $(B_{JJ})^{-1}$ yields

$$(B_{JJ})^{-1}B_J(x^{r+1}-x^r)+(B_{JJ})^{-1}M_J(x^r-x^*)=0, \qquad r=t,\cdots,s-1.$$

This in turn implies, after some rearrangement of terms, that

$$\begin{aligned} x_J^{r+1}-x_J^* &= (I-(B_{JJ})^{-1}M_{JJ})(x_J^r-x_J^*)-(B_{JJ})^{-1}M_{J\tilde{J}}(x_{\tilde{J}}^r-x_{\tilde{J}}^*) \\ &\quad -(B_{JJ})^{-1}B_{J\tilde{J}}(x_{\tilde{J}}^{r+1}-x_{\tilde{J}}^r), \qquad r=t,\cdots,s-1. \end{aligned}$$

By letting $G=I-(B_{JJ})^{-1}M_{JJ}$ and successively applying the above recursion for $r=t,\cdots,s-1$, we obtain

$$\begin{aligned} x_J^s-x_J^* &= (G)^h(x_J^t-x_J^*)-\sum_{k=0}^{h-1}(G)^{h-k-1}(B_{JJ})^{-1}M_{J\tilde{J}}(x_{\tilde{J}}^{t+k}-x_{\tilde{J}}^*) \\ &\quad -\sum_{k=0}^{h-1}(G)^{h-k-1}(B_{JJ})^{-1}B_{J\tilde{J}}(x_{\tilde{J}}^{t+k+1}-x_{\tilde{J}}^{t+k}), \end{aligned} \tag{3.4}$$

where we denote $h=s-t$. Now we estimate the last sum in (3.4). Let

$$y^k=(B_{JJ})^{-1}B_{J\tilde{J}}(x_{\tilde{J}}^{t+k}-x_{\tilde{J}}^*), \qquad k=0,1,\cdots,h.$$

Then the last sum in (3.4) can be rewritten as $\sum_{k=0}^{h-1}(G)^{h-k-1}(y^{k+1}-y^k)$. By rearranging the terms within the summation sign, we obtain an alternative form for this sum:

$$\begin{aligned} \sum_{k=0}^{h-1}(G)^{h-k-1}(y^{k+1}-y^k) &= \sum_{k=0}^{h-1}(G)^{h-k-1}y^{k+1}-\sum_{k=0}^{h-1}(G)^{h-k-1}y^k \\ &= \sum_{k=1}^{h-1}(G)^{h-k-1}(G)y^k+y^h-(G)^{h-1}y^0-\sum_{k=1}^{h-1}(G)^{h-k-1}y^k \\ &= \sum_{k=1}^{h-1}(G)^{h-k-1}(G-I)y^k+y^h-(G)^{h-1}y^0. \end{aligned}$$

Since $G-I=-(B_{JJ})^{-1}M_{JJ}$, this together with (3.4) implies that

$$\begin{aligned} x_J^s-x_J^* &= (G)^h(x_J^t-x_J^*)-\sum_{k=0}^{h-1}(G)^{h-k-1}(B_{JJ})^{-1}M_{J\tilde{J}}(x_{\tilde{J}}^{t+k}-x_{\tilde{J}}^*) \\ &\quad +\sum_{k=1}^{h-1}(G)^{h-k-1}(B_{JJ})^{-1}M_{JJ}y^k-y^h+(G)^{h-1}y^0. \end{aligned}$$

Let $H=I-M_{JJ}(B_{JJ})^{-1}$. Then $G=(B_{JJ})^{-1}HB_{JJ}$, so that $(G)^{h-k-1}=(B_{JJ})^{-1}(H)^{h-k-1}B_{JJ}$ for all k. This together with the above equation yields

$$\begin{aligned} x_J^s-x_J^* &= (G)^h(x_J^t-x_J^*)-\sum_{k=0}^{h-1}(B_{JJ})^{-1}(H)^{h-k-1}M_{J\tilde{J}}(x_{\tilde{J}}^{t+k}-x_{\tilde{J}}^*) \\ &\quad +\sum_{k=1}^{h-1}(B_{JJ})^{-1}(H)^{h-k-1}M_{JJ}y^k-y^h+(G)^{h-1}y^0. \end{aligned}$$

Also, since $J\subseteq I^*$, we have from Lemma 4 that $\|(H)^{h-k-1}z\|\leqq\tau_J(\rho_J)^{h-k-1}\|z\|$ for any $z\in\operatorname{Span}(M_{JJ})$ and $\|(G)^hz\|\leqq\Delta\|z\|$ for any z. This, together with the above equation and the fact Span $(M_{J\tilde{J}}) \subseteq$ Span (M_{JJ}) (cf. Lemma 5), implies

$$\begin{aligned}
\|x^s_J - x^*_J\| &\leqq \|(G)^b(x^t_J - x^*_J)\| + \sum_{k=0}^{h-1} \|(B_{JJ})^{-1}\| \, \|(H)^{h-k-1} M_{J\tilde{J}}(x^{t+k}_{\tilde{J}} - x^*_{\tilde{J}})\| \\
&\quad + \sum_{k=1}^{h-1} \|(B_{JJ})^{-1}\| \, \|(H)^{h-k-1} M_{JJ} y^k\| + \|y^h\| + \|(G)^{h-1} y^0\| \\
&\leqq \Delta \|x^t_J - x^*_J\| + \sum_{k=0}^{h-1} \|(B_{JJ})^{-1}\| \tau_J (\rho_J)^{h-k-1} \|M_{J\tilde{J}}(x^{t+k}_{\tilde{J}} - x^*_{\tilde{J}})\| \\
&\quad + \sum_{k=1}^{h-1} \|(B_{JJ})^{-1}\| \tau_J (\rho_J)^{h-k-1} \|M_{JJ} y^k\| + \|y^h\| + \Delta \|y^0\| \\
&\leqq \Delta \|x^t_J - x^*_J\| + \tau_J \|(B_{JJ})^{-1}\| \, \|M_{J\tilde{J}}\| \sum_{k=0}^{h-1} (\rho_J)^{h-k-1} \max_{r \in \{t, \cdots, s-1\}} \|x^r_{\tilde{J}} - x^*_{\tilde{J}}\| \\
&\quad + \tau_J \|(B_{JJ})^{-1}\| \, \|M_{JJ}\| \sum_{k=1}^{h-1} (\rho_J)^{h-k-1} \max_{k \in \{1, \cdots, h-1\}} \|y^k\| + \|y^h\| + \Delta \|y^0\| \\
&\leqq \Delta \|x^t_J - x^*_J\| + \tau_J \|(B_{JJ})^{-1}\| \, \|M_{J\tilde{J}}\| (1 - \rho_J)^{-1} \max_{r \in \{t, \cdots, s-1\}} \|x^t_{\tilde{J}} - x^*_{\tilde{J}}\| \\
&\quad + \tau_J \|(B_{JJ})^{-1}\| \, \|M_{JJ}\| (1 - \rho_J)^{-1} \max_{k \in \{1, \cdots, h-1\}} \|y^k\| + \|y^h\| + \Delta \|y^0\|.
\end{aligned}$$

Since $y^k = (B_{JJ})^{-1} B_{J\tilde{J}}(x^{t+k}_{\tilde{J}} - x^*_{\tilde{J}})$, we also have $\|y^k\| \leqq \|(B_{JJ})^{-1} B_{J\tilde{J}}\| \, \|x^{t+k}_{\tilde{J}} - x^*_{\tilde{J}}\|$ for every k, and the lemma is proved. □

By using Lemmas 6–8, we can now prove our main result that $\{x^r\}$ converges (see Theorem 1). The basic idea of the proof is to show that those coordinates of x^r that are bounded sufficiently far away from the boundary of X are essentially unaffected by the rest. This then allows us to treat these coordinates as if they are unconstrained and by using the contraction property of $\mathcal{A}_B$ on them, we conclude convergence for these coordinates.

In our proof, we will make frequent use of the following scalars:

$$\begin{aligned}
&\sigma_0 = 1, \\
&\sigma_k = \Delta + 3 + \beta + (\beta + 1)\sigma_{k-1}, \qquad k = 1, 2, \cdots, n.
\end{aligned}$$

(Note that $\sigma_k \geqq 1$ for all k and is monotonically increasing with k.) We will also use the fact (cf. Lemmas 6(a) and 7(b)) that, for any $\delta > 0$, there exists a scalar r_δ such that

$$\phi(x^r) \leqq \delta \quad \forall r \geqq r_\delta, \tag{3.5}$$

$$\|x^{r+1} - x^r\| \leqq \delta \quad \forall r \geqq r_\delta. \tag{3.6}$$

Since $x^*_i = 0$ for all $i \notin I^*$ and all $x^* \in X^*$ (cf. (2.11)), it immediately follows from (3.5) that

$$x^r_i \leqq \delta \quad \forall r \geqq r_\delta, \quad \forall i \notin I^*. \tag{3.7}$$

Our proof comprises a sequence of three lemmas. The first lemma roughly shows that if those coordinates of x^r which start near the boundary of X stay near the boundary while the remaining coordinates start far from the boundary, then x^r will stay close to the optimal solution set.

LEMMA 9. *Fix any $\delta > 0$ and let r_δ be a scalar such that* (3.5)-(3.6) *hold. If for some $k \in \{1, \cdots, n\}$, some nonempty $J \subseteq I^*$ and some two integers $t' > t \geqq r_\delta$ we have*

$$x_i^t > \sigma_k \delta \quad \forall i \in J, \tag{3.8}$$

$$x_i^r \leqq \sigma_{k-1} \delta \quad \forall i \notin J, \quad \forall r = t, t+1, \cdots, t'-1, \tag{3.9}$$

then the following hold:

(a) $x_i^{t'} > \sigma_{k-1}\delta$, *for all* $i \in J$.

(b) *There exists an $x^* \in X^*$ such that*

$$\|x^r - x^*\|_\infty \leqq \sigma_k \delta \quad \forall r = t, t+1, \cdots, t'-1.$$

Proof. Let x^* be any element of X^* satisfying $\phi(x^t) = \|x^t - x^*\|$. Then we have from (3.5) that

$$\|x^t - x^*\| \leqq \delta. \tag{3.10}$$

Also we have from (3.9) that, for all $i \notin J$, $x_i^* \leqq x_i^t + \|x^t - x^*\| \leqq \sigma_{k-1}\delta + \|x^t - x^*\|$, which together with (3.10) implies $0 \leqq x_i^* \leqq \sigma_{k-1}\delta + \delta$. Since $0 \leqq x_i^r \leqq \sigma_{k-1}\delta$ for $r = t, t+1, \cdots, t'-1$ (cf. (3.9)), this in turn implies that

$$|x_i^r - x_i^*| \leqq \sigma_{k-1}\delta + \delta \quad \forall i \notin J, \quad \forall r = t, t+1, \cdots, t'-1. \tag{3.11}$$

Next we prove by induction that, for $r = t, t+1, \cdots, t'-1$,

$$x_i^r > \sigma_{k-1}\delta + \delta \quad \forall i \in J. \tag{3.12}$$

Equation (3.12) clearly holds for $r = t$ (cf. (3.8) and $\sigma_k \geqq \sigma_{k-1} + 1$). Suppose that (3.12) holds for $r = t, t+1, \cdots, s$, for some $s \in \{t, t+1, \cdots, t'-2\}$. We will prove that it also holds for $r = s+1$. Since $x_i^r > 0$ for all $i \in J$ and all $r = t+1, \cdots, s$ (cf. (3.12)), we have from Lemma 8 that

$$\|x_J^s - x_J^*\| \leqq \Delta \|x_J^t - x_J^*\| + \beta \max_{r \in \{t, \cdots, s\}} \|x_{\bar{J}}^r - x_{\bar{J}}^*\|_\infty,$$

which together with (3.10) and (3.11) implies

$$\|x_J^s - x_J^*\| \leqq \Delta\delta + \beta(\sigma_{k-1}\delta + \delta). \tag{3.13}$$

Then we have that, for any $i \in J$,

$$\begin{aligned} x_i^{s+1} &\geqq x_i^t - \|x_J^t - x_J^{s+1}\| \\ &\geqq x_i^t - (\|x_J^t - x_J^*\| + \|x_J^* - x_J^s\| + \|x_J^s - x_J^{s+1}\|) \\ &> \sigma_k\delta - (\delta + \|x_J^* - x_J^s\| + \delta) \\ &\geqq \sigma_k\delta - (\delta + (\Delta\delta + \beta\sigma_{k-1}\delta + \beta\delta) + \delta) \\ &= \sigma_{k-1}\delta + \delta, \end{aligned}$$

where the strict inequality follows from (3.6), (3.8), and (3.10). This completes the induction and proves that (3.12) holds for $r = t, t+1, \cdots, t'-1$. Since (3.12) holds for $r = t, t+1, \cdots, t'-1$, it can be seen from the argument above that (3.13) holds for $s = t, t+1, \cdots, t'-1$, which when combined with (3.11) (and using the facts $\beta > 1$ and $\|z\|_\infty \leqq \|z\|$ for all z) yields

$$\|x^r - x^*\|_\infty \leqq (\Delta + \beta\sigma_{k-1} + \beta)\delta \quad \forall r = t, t+1, \cdots, t'-1.$$

Since $\Delta+\beta\sigma_{k-1}+\beta\leqq\sigma_k$, this proves part (b). From (3.12) with $r=t'-1$, we have that, for all $i\in J$,

$$\begin{aligned} x_i^{t'} &\geqq x_i^{t'-1}-\|x^{t'-1}-x^{t'}\| \\ &>\sigma_{k-1}\delta+\delta-\|x^{t'-1}-x^{t'}\|. \end{aligned}$$

Since $\|x^{t'-1}-x^{t'}\|\leqq\delta$ (cf. (3.6)), this proves part (a). □

The following lemma extends Lemma 9 by removing the assumption that the coordinates that start near the boundary of X remain near the boundary (while still assuming that the remaining coordinates start far from the boundary).

LEMMA 10. *Fix any $\delta>0$ and let r_δ be a scalar such that* (3.5)-(3.6) *hold. If for some $k\in\{1,\cdots,n\}$, some $J\subseteq I^*$ with* Card $(J)\geqq$ Card $(I^*)-k+1$ *and some integer $t\geqq r_\delta$ we have*

$$x_i^t>\sigma_k\delta\quad\forall i\in J,\qquad x_i^t\leqq\sigma_{k-1}\delta\quad\forall i\notin J;$$

then there exist an $x^*\in X^*$ and a $\bar{t}\geqq t$ satisfying

$$\|x^r-x^*\|_\infty\leqq\sigma_k\delta\quad\forall r\geqq\bar{t}. \tag{3.14}$$

Proof. Our proof is by induction on k. By Lemma 9(b), we see that the claim holds for $k=1$. (Since in this case $J=I^*$, then, by (3.7), condition (3.9) is satisfied for all $t'\geqq t$, so Lemma 9(b) yields that there exists an $x^*\in X^*$ such that $\|x^r-x^*\|_\infty\leqq\sigma_k\delta$ for all $r\geqq t$.) Suppose that the claim holds for $k=1,2,\cdots,h-1$, for some $h\geqq2$. We show below that it also holds for $k=h$.

Fix any $J\subseteq I^*$ with Card $(J)\geqq$ Card $(I^*)-h+1$ and any integer t for which

$$x_i^t>\sigma_h\delta\quad\forall i\in J, \tag{3.15}$$

$$x_i^t\leqq\sigma_{h-1}\delta\quad\forall i\notin J. \tag{3.16}$$

We consider two cases.

Case 1. $x_i^r\leqq\sigma_{h-1}\delta$, for all $i\notin J$ and all $r\geqq t$. Since $x_i^t>\sigma_h\delta$, for all $i\in J$ (cf. (3.15)), it immediately follows from Lemma 9(b) that there exists an $x^*\in X^*$ such that

$$\|x^r-x^*\|_\infty\leqq\sigma_h\delta\quad\forall r\geqq t.$$

This shows that (3.14) holds for $k=h$ (with $\bar{t}=t$ and with the above choice of x^*).

Case 2. There exists an $r>t$ and an $i\notin J$ such that $x_i^r>\sigma_{h-1}\delta$. Let

$$t'=\text{smallest } r\ (r>t) \text{ such that } x_i^r>\sigma_{h-1}\delta \text{ for some } i\notin J.$$

Then, by (3.16), $x_i^r\leqq\sigma_{h-1}\delta$ for all $i\notin J$ and all $r=t,t+1,\cdots,t'-1$. Since $x_i^t>\sigma_h\delta$, for all $i\in J$ (cf. (3.15)), Lemma 9(a) yields that

$$x_i^{t'}>\sigma_{h-1}\delta\quad\forall i\in J. \tag{3.17}$$

Consider the $h+1$ intervals

$$[0,\sigma_0\delta],\quad(\sigma_0\delta,\sigma_1\delta],\quad(\sigma_1\delta,\sigma_2\delta],\quad\cdots,\quad(\sigma_{h-2}\delta,\sigma_{h-1}\delta],\quad(\sigma_{h-1}\delta,\infty). \tag{3.18}$$

We have from (3.17) and the fact $x_i^{t'}>\sigma_{h-1}\delta$ for some $i\notin J$ that the $(h+1)$st interval contains at least Card $(J)+1$ elements from the set $\{x_1^{t'},x_2^{t'},\cdots,x_n^{t'}\}$. Also, (3.7) and $\sigma_0=1$ imply that the first interval contains at least $n-$Card (I^*) elements from the same set. Since Card $(J)\geqq$ Card $(I^*)-h+1$, this leaves at most $h-2$ elements from that set to go into the remaining $h-1$ intervals. Hence, by the pigeon hole principle, there must exist some $j\in\{1,2,\cdots,h-1\}$ such that

$$x_i^{t'}\notin(\sigma_{j-1}\delta,\sigma_j\delta]\quad\forall i.$$

Let h' be the largest j for which this occurs. Then the interval $(\sigma_{h'}\delta, \infty)$ contains at least Card $(J)+h-h'$ elements from the set $\{x_1^{t'}, x_2^{t'}, \cdots, x_n^{t'}\}$. Let J' be the index set for these elements, i.e., $J'=\{i \mid x_i^{t'} > \sigma_{h'}\delta\}$. Then we have

$$x_i^{t'} > \sigma_{h'}\delta \quad \forall i \in J', \qquad x_i^{t'} \leqq \sigma_{h'-1}\delta \quad \forall i \notin J',$$

and

$$\begin{aligned} \text{Card}\,(J') &\geqq \text{Card}\,(J)+h-h' \\ &\geqq \text{Card}\,(I^*)+1-h'. \end{aligned}$$

Moreover, by (3.7), we have $J' \subseteq I^*$. Since $h' < h$, we can apply our induction hypothesis to h', J', and t' to conclude that there exists an $x^* \in X^*$ and a $\bar{t} \geqq t'$ satisfying

$$\|x^r - x^*\|_\infty \leqq \sigma_{h'}\delta \quad \forall r \geqq \bar{t}.$$

Since $\sigma_{h'} \leqq \sigma_h$, this shows that (3.14) holds for $k=h$ (with the given $\bar{t}$ and x^*).

Hence in either case the claim holds for $k=h$. This then completes the induction on k and proves the lemma. □

We are now ready to prove the following key lemma.

LEMMA 11. *For any $\delta > 0$, there exists an $x^* \in X^*$ and an $\hat{r} > 0$ such that*

$$\|x^r - x^*\|_\infty \leqq \sigma_n\delta + \delta \quad \forall r \geqq \hat{r}. \tag{3.19}$$

Proof. The proof is based on Lemma 10. Fix any $\delta > 0$ and let r_δ be a scalar such that (3.5)-(3.6) hold. Choose any integer $\bar{r} \geqq r_\delta$ and consider the two possible cases: either (i) $x_i^r \leqq \sigma_n\delta$ for all i and all $r \geqq \bar{r}$, or (ii) there exists a $t \geqq \bar{r}$ and an i such that $x_i^t > \sigma_n\delta$. In case (i), let x^* be an element of X^* such that $\phi(x^{\bar{r}}) = \|x^{\bar{r}} - x^*\|$. Then we have from (3.5) that, for all i,

$$0 \leqq x_i^* \leqq x_i^{\bar{r}} + \|x^{\bar{r}} - x^*\| \leqq \sigma_n\delta + \delta.$$

Since $0 \leqq x_i^r \leqq \sigma_n\delta$ for all i and all $r \geqq \bar{r}$, this implies that

$$\|x^r - x^*\|_\infty \leqq \sigma_n\delta + \delta \quad \forall r \geqq \bar{r}.$$

Hence (3.19) holds with $\hat{r} = \bar{r}$ and with the above choice of x^*. Now consider case (ii). In this case, by the pigeon hole principle, one of the following n intervals

$$(\sigma_0\delta, \sigma_1\delta], \quad (\sigma_1\delta, \sigma_2\delta], \quad \cdots, \quad (\sigma_{n-1}\delta, \sigma_n\delta]$$

does not contain any element from $\{x_1^t, x_2^t, \cdots, x_n^t\}$, i.e., there exists $j \in \{1, 2, \cdots, n\}$ such that

$$x_i^t \notin (\sigma_{j-1}\delta, \sigma_j\delta] \quad \forall i.$$

Choose k to be the *largest* such j and let $J = \{i \mid x_i^t > \sigma_k\delta\}$. Then Card $(J) \geqq n-k+1$ and

$$x_i^t > \sigma_k\delta \quad \forall i \in J, \qquad x_i^t \leqq \sigma_{k-1}\delta \quad \forall i \notin J.$$

Moreover, by (3.7) and $\sigma_k \geqq 1$, we see that $J \subseteq I^*$. Hence the assumptions of Lemma 10 are satisfied by k, J, and t, and it follows from Lemma 10 that there exists an $x^* \in X^*$ and a $\bar{t} \geqq t$ satisfying

$$\|x^r - x^*\|_\infty \leqq \sigma_k\delta \quad \forall r \geqq \bar{t}.$$

Since $\sigma_k \leqq \sigma_n$, this shows that (3.19) holds (with the given x^* and with $\hat{r} = \bar{t}$). □

The following main convergence result then follows as a corollary of Lemma 11.

THEOREM 1. *The matrix splitting algorithm* (1.4)-(1.5) *is well defined and it generates a sequence of iterates* $\{x^r\}$ *converging to an element of* X^*.

Proof. The algorithm is well defined by Lemma 2(a). Now, for any $\varepsilon > 0$, Lemma 11 shows that there exists an $x^* \in X^*$ and an $\hat{r} > 0$ such that

$$\|x^r - x^*\|_\infty < \varepsilon/2 \quad \forall r \geqq \hat{r}.$$

Hence, for all $r_1, r_2 > \hat{r}$, there holds

$$\begin{aligned}\|x^{r_1} - x^{r_2}\|_\infty &\leqq \|x^{r_1} - x^*\|_\infty + \|x^* - x^{r_2}\|_\infty \\ &\leqq \varepsilon/2 + \varepsilon/2 = \varepsilon.\end{aligned}$$

This implies that $\{x^r\}$ is a Cauchy sequence so that it converges. By Lemma 7(b), it converges to an element of X^*. □

4. SOR matrix splitting algorithms. In this section we consider three extensions of the basic algorithm (1.4)-(1.5). First, we consider one that adds an under/overrelaxation parameter to the algorithm. This extension is motivated by the block SOR methods of Cottle, Golub, and Sacher [CGS78], of Cottle and Goheen [CoG78], and of Cottle and Pang [CoP82] (which introduced a mechanism for overrelaxation) and the methods of Mangasarian [Man77] and of Aganagic [Aga78] (which introduced a mechanism for underrelaxation). Second, we consider a Gauss-Seidel extension of the basic algorithm. In this algorithm, only a subset of the coordinates are relaxed at each iteration while the other coordinates are held fixed. Last, we consider an SOR extension of the basic algorithm which allows noncyclic order of relaxation. This third algorithm contains the previous two as special cases but is shown to be convergent only in a certain weak sense.

We first describe the under/overrelaxation extension. In the algorithm, we choose a splitting (B, C) of M and a relaxation parameter $\bar{\omega}$ satisfying

$$0 < \bar{\omega}, \qquad B - C + (1 - \bar{\omega})M \text{ is positive definite.} \tag{4.1}$$

We also choose a second relaxation parameter $\underline{\omega}$ satisfying

$$0 < \underline{\omega} \leqq \min\{1, \bar{\omega}\}, \tag{4.2}$$

and an $n \times n$ positive *diagonal* matrix D. Then, for any chosen $x^0 \in X$, we generate a sequence of vectors $\{x^0, x^1, \cdots\}$ in X by the formula

$$x^{r+1} = (1 - \omega^r)x^r + \omega^r \hat{x}^r, \tag{4.3}$$

where $\hat{x}^r$ is a solution of the equation

$$y = [y - D(By + Cx^r + q)]^+, \tag{4.4}$$

and ω^r is *any* scalar in $[\underline{\omega}, \bar{\omega}]$ such that x^{r+1} given by (4.3) is in X.

Note that if (B, C) is a regular splitting and

$$0 < \bar{\omega} < 1 + \frac{1}{\|Q^{-1/2}MQ^{-1/2}\|},$$

where Q denotes the symmetric part of $B - C$, i.e., $Q = ((B - C) + (B - C)^T)/2$, then (4.1) is satisfied. Hence, the algorithm above contains as a special case the algorithm (1.4)-(1.5) (let $\underline{\omega} = \bar{\omega} = 1$ and D be the $n \times n$ identity matrix). The relaxation parameter we introduced in (4.3) is useful mainly when $\omega^r > 1$ (i.e., *overrelaxation* [OrR70]), which in some cases can significantly improve the convergence. Nonetheless, the case of *underrelaxation*, i.e., $\omega^r < 1$, is also of some practical interest since, in this case, it is only required that $B-C$ be positive definite on the null space of M (instead of on the entire space) in order for (4.1) to hold. The purpose for introducing the matrix D in (4.4) is, from the point of view of convergence, largely cosmetic as the presence of D does not change the sequence of iterates generated. (To see this, note that since D is a diagonal matrix, y is a solution of (4.4) if and only if $y \in X$ and $y_i = 0$ for all i such that $D_{ii}(B_i y + C_i x^r + q_i) > 0$. Since $D_{ii} > 0$ for all i, this set of conditions is equivalent to $y \in X$ and $y_i = 0$ for all i such that $B_i y + C_i x^r + q_i > 0$, which in turn is equivalent to $y = [y - (By + Cx^r + q)]^+$ or, by (1.4), $y = \mathcal{A}_B(x^r)$.) However, by choosing D to match the structure of B and C, we can in some cases simplify the form of the iteration (see § 5 for examples). Note that since the sequence of iterates generated is independent of D, we can also allow D to be time-varying.

By modifying the argument used in §§ 2 and 3, we can show that the algorithm (4.1)-(4.4) is convergent.

THEOREM 2. *For any splitting* (B, C) *of* M *and any scalars* $\underline{\omega}$, $\bar{\omega}$ *satisfying* (4.1)-(4.2), *any* $n \times n$ *positive diagonal matrix* D *and any* $x^0 \in X$, *the sequence of iterates* $\{x^r\}$ *generated by* (4.3)-(4.4) *is well defined and converges to an element of* X^*.

Sketch of proof. Since $2B = (B - C + (1-\bar{\omega})M) + \bar{\omega}M$ (cf. $M = B + C$) and $\bar{\omega}M$ is positive semidefinite, we have from (4.1) that B is positive definite. The proof of Lemma 2(a) then shows that $\mathcal{A}_B$ is a well-defined point-to-point mapping. For $r = 0, 1, \cdots$, let $\hat{x}^r$ be a solution of (4.4). Then from the preceding discussion we have that

$$\hat{x}^r = \mathcal{A}_B(x^r), \qquad r = 0, 1, \cdots,$$

so that $\{\hat{x}^r\}$ is well defined. Since $x^{r+1} = (1-\omega^r)x^r + \omega^r \hat{x}^r$ for all r (cf. (4.3)), $\{x^r\}$ is well defined.

It remains to show that $\{x^r\}$ is convergent. The proof of this is analogous to that of Theorem 1, with Lemmas 4, 6, and 8 replaced by more general versions of themselves that take into account the relaxation parameters. More precisely, by applying Lemma 3 with the identifications $Q \leftrightarrow M_{JJ}$, $N \leftrightarrow B_{JJ}/\omega$, $\underline{\theta} \leftrightarrow \underline{\omega}/\bar{\omega}$, it is easily seen that the following generalization of Lemma 4 holds:

(a) For any nonempty $J \subseteq \{1, \cdots, n\}$, there exist $\rho_J \in (0, 1)$ and $\tau_J > 0$ such that

$$\left\| \prod_{h=s}^{s+k-1} (I - \omega^h M_{JJ}(B_{JJ})^{-1}) z \right\| \leqq \tau_J (\rho_J)^k \|z\| \quad \forall k \geqq 1, \quad \forall s \geqq 0, \quad \forall z \in \operatorname{Span}(M_{JJ}).$$

(b) There exists a $\Delta > 0$ such that, for any nonempty $J \subseteq \{1, \cdots, n\}$,

$$\left\| \prod_{h=s}^{s+k-1} (I - \omega^h (B_{JJ})^{-1} M_{JJ}) z \right\| \leqq \Delta \|z\| \quad \forall k \geqq 1, \quad \forall s \geqq 0, \quad \forall z.$$

By using the above generalized version of Lemma 4 in place of Lemma 4, it can be verified that Lemma 8, with "$x_i^r > 0$ for all $i \in J$ and all $r = t+1, \cdots, s$" replaced by "$\hat{x}_i^r > 0$ for all $i \in J$ and all $r = t+1, \cdots, s$," still holds (possibly with a different constant β that depends not only on M and B, but also on $\underline{\omega}$ and $\bar{\omega}$). Also, it can be verified by using (4.1), (4.3), (4.4), and $\omega^r \in [\underline{\omega}, \bar{\omega}]$, for all r, that relation (3.1) still holds (with a slightly different γ given by $\gamma =$ smallest eigenvalue of the symmetric part of $((1-\bar{\omega})M + B - C)/\underline{\omega})$, so it readily follows that Lemmas 6 and 7 also hold.

Now, we have from (4.3) that, for all r, $x^{r+1} - \hat{x}^r = (1 - 1/\omega^r)(x^{r+1} - x^r)$, so

$$\|x^{r+1} - \hat{x}^r\| \leqq \max\{1, 1/\underline{\omega}\} \|x^{r+1} - x^r\|.$$

Then, by redefining the scalar σ_0 to be $1 + \max\{1, 1/\underline{\omega}\}$ with the other scalars $\sigma_1, \cdots, \sigma_n$ recursively defined as before, the proof of Lemma 9 (which depends on Lemmas 6-8 only) still goes through. Lemmas 10 and 11 then follow from Lemmas 6, 7, and 9 as before. □

Remark 5. We can also use *different relaxation parameters for different coordinates* provided that the relaxation parameters are *fixed.* More precisely, let us consider the following underrelaxed algorithm:

$$x^{r+1}=\begin{bmatrix}1-\bar{\omega}_1 & & \\ & \ddots & \\ & & 1-\bar{\omega}_n\end{bmatrix}x^r+\begin{bmatrix}\bar{\omega}_1 & & \\ & \ddots & \\ & & \bar{\omega}_n\end{bmatrix}\mathscr{A}_B(x^r), \qquad r=0,1,\cdots,$$

where $x^0\in X$, $\bar{\omega}_1,\cdots,\bar{\omega}_n$ are fixed scalars in the interval $(0,1]$, and (B, C) is a splitting of M for which the matrix

$$2B\begin{bmatrix}1/\bar{\omega}_1 & & \\ & \ddots & \\ & & 1/\bar{\omega}_n\end{bmatrix}-M$$

is positive definite. In the special case where $\bar{\omega}_1=\cdots=\bar{\omega}_n$, this algorithm reduces to the algorithm (4.1)-(4.4) using fixed underrelaxation. By suitably modifying the proof of Theorem 2, it can be shown that this underrelaxed algorithm is convergent.

Now we consider a Gauss-Seidel type algorithm. Let $\{1,\cdots,n\}$ be partitioned into m nonempty, mutually disjoint subsets $I_1, I_2,\cdots,I_m$ (i.e., $I_i\cap I_j=\varnothing$ if $i\neq j$ and $I_1\cup\cdots\cup I_m=\{1,\cdots,n\}$). For $j=1,\cdots,m$, we choose a regular splitting $(B_{I_jI_j}, C_{I_jI_j})$ of $M_{I_jI_j}$ and define a corresponding mapping $\mathscr{A}_j: X\mapsto X$ by

$$\text{(4.5)}\qquad \mathscr{A}_j(x)=\{y\in\Re^n\,|\,y_{I_j}=[y_{I_j}-(B_{I_jI_j}y_{I_j}+C_{I_jI_j}x_{I_j}+M_{I_j\tilde{I}_j}x_{\tilde{I}_j}+q_{I_j})]_j^+,\ y_{\tilde{I}_j}=x_{\tilde{I}_j}\},$$

where $[\cdot]_j^+$ denotes the orthogonal projection onto the box $[0,\infty)^{\mathrm{Card}(I_j)}$. By Lemma 2(a), $\mathscr{A}_j$ is a well-defined point-to-point mapping. The mapping $\mathscr{A}_j$ has the effect of applying a matrix splitting iteration to the subset of coordinates indexed by I_j, while the other coordinates are held fixed. The Gauss-Seidel matrix splitting (GS-MS) algorithm generates a sequence of iterates by applying cyclically the mappings $\mathscr{A}_1,\cdots,\mathscr{A}_m$:

GS-MS ALGORITHM. Choose an $x^0\in X$. Generate a sequence of vectors $\{x^0, x^1,\cdots\}$ in X by the formula

$$\text{(4.6)}\qquad x^{r+1}=(\mathscr{A}_m\circ\cdots\circ\mathscr{A}_2\circ\mathscr{A}_1)(x^r), \qquad r=0,1,\cdots.$$

It is easily seen that in the special case where $m=1$, this algorithm reduces to the algorithm (4.1)-(4.4) with relaxation parameters $\underline{\omega}=\bar{\omega}=1$.

By extending the proof of Theorem 2, we can show that the GS-MS algorithm is convergent.

THEOREM 3. *The sequence of iterates generated by the* GS-MS *algorithm* (4.5)-(4.6) *converges to an element of* X^*.

Sketch of proof. Similar to the proof of Theorem 2, it suffices to show that Lemmas 6-8 still hold.

We first show that Lemmas 6 and 7 still hold. Since the I_j's are disjoint, we have from (4.6) that

$$\text{(4.7)}\qquad x_{I_j}^{r+1}=\mathscr{A}_j(x_{I_1}^{r+1},\cdots,x_{I_j}^{r+1},x_{I_{j+1}}^r,\cdots,x_{I_m}^r), \qquad j=1,\cdots,m, \quad \forall r.$$

Then, by using (4.5) and an argument analogous to that for (3.1), we obtain that

$$f(x_{I_1}^{r+1},\cdots,x_{I_j}^{r+1},x_{I_{j+1}}^r,\cdots,x_{I_m}^r)\leqq f(x_{I_1}^{r+1},\cdots,x_{I_{j-1}}^{r+1},x_{I_j}^r,\cdots,x_{I_m}^r)-\gamma\|x_{I_j}^{r+1}-x_{I_j}^r\|^2,$$

for all r and all j, where $\gamma > 0$ denotes the smallest eigenvalue of the symmetric part of $B_{I_jI_j} - C_{I_jI_j}$, minimized over all j. By applying the above inequality recursively for all j, we obtain that

$$f(x^{r+1}) \leqq f(x^r) - \gamma \sum_{j=1}^{m} \|x_{I_j}^{r+1} - x_{I_j}^r\|^2 = f(x^r) - \gamma \|x^{r+1} - x^r\|^2 \quad \forall r,$$

and it readily follows that Lemmas 6 and 7 hold.

Now we show that Lemma 8 still holds (possibly with a different β). Consider any $J \subseteq I^*$ and suppose that for some two integers $s \geqq t \geqq 0$ we have $x_i^r > 0$ for all $i \in J$ and all $r = t+1, t+2, \cdots, s$. Let $J_j = J \cap I_j$ and $\hat{J}_j = \tilde{J} \cap I_j$. Then $I_j = J_j \cup \hat{J}_j$ for all j, and we have from (4.5) and (4.7) that, for any $r \in \{t, \cdots, s-1\}$, there holds

$$\begin{aligned} 0 &= B_{J_jI_j} x_{I_j}^{r+1} + C_{J_jI_j} x_{I_j}^r + \sum_{k<j} M_{J_jI_k} x_{I_k}^{r+1} + \sum_{k>j} M_{J_jI_k} x_{I_k}^r + q_{J_j} \\ &= B_{J_jJ_j} x_{J_j}^{r+1} + C_{J_jJ_j} x_{J_j}^r + \sum_{k<j} M_{J_jJ_k} x_{J_k}^{r+1} + \sum_{k>j} M_{J_jJ_k} x_{J_k}^r \\ &\qquad + B_{J_j\hat{J}_j} x_{\hat{J}_j}^{r+1} + C_{J_j\hat{J}_j} x_{\hat{J}_j}^r + \sum_{k<j} M_{J_j\hat{J}_k} x_{\hat{J}_k}^{r+1} + \sum_{k>j} M_{J_j\hat{J}_k} x_{\hat{J}_k}^r + q_{J_j}, \qquad j = 1, \cdots, m. \end{aligned}$$

By rewriting the fifth to the eighth terms in the above expression as

$$B_{J_j\hat{J}_j}(x_{\hat{J}_j}^{r+1} - x_{\hat{J}_j}^r) + \sum_{k<j} M_{J_j\hat{J}_k}(x_{\hat{J}_k}^{r+1} - x_{\hat{J}_k}^r) + \sum_{k=1}^{m} M_{J_j\hat{J}_k} x_{\hat{J}_k}^r,$$

we can express the above set of equations using a single matrix splitting as follows:

$$\begin{aligned} &\begin{bmatrix} B_{J_1J_1} & & & & & \\ M_{J_2J_1} & B_{J_2J_2} & & & & \\ \vdots & \ddots & \ddots & & & \\ M_{J_jJ_1} & \cdots & M_{J_jJ_{j-1}} & B_{J_jJ_j} & & \\ \vdots & & & \ddots & \ddots & \\ M_{J_mJ_1} & & \cdots & & M_{J_mJ_{m-1}} & B_{J_mJ_m} \end{bmatrix} x_J^{r+1} \\ &\quad + \begin{bmatrix} C_{J_1J_1} & M_{J_1J_2} & & \cdots & & M_{J_1J_m} \\ & \ddots & \ddots & & & \vdots \\ & & C_{J_iJ_i} & M_{J_jJ_{j+1}} & \cdots & M_{J_jJ_m} \\ & & & \ddots & \ddots & \vdots \\ & & & & C_{J_{m-1}J_{m-1}} & M_{J_{m-1}J_m} \\ & & & & & C_{J_mJ_m} \end{bmatrix} x_J^r \\ &\quad + \begin{bmatrix} B_{J_1\hat{J}_1} & & & & & \\ M_{J_2\hat{J}_1} & B_{J_2\hat{J}_2} & & & & \\ \vdots & \ddots & \ddots & & & \\ M_{J_j\hat{J}_1} & \cdots & M_{J_j\hat{J}_{j-1}} & B_{J_j\hat{J}_j} & & \\ \vdots & & & \ddots & \ddots & \\ M_{J_m\hat{J}_1} & & \cdots & & M_{J_m\hat{J}_{m-1}} & B_{J_m\hat{J}_m} \end{bmatrix} \cdot (x_{\tilde{J}}^{r+1} - x_{\tilde{J}}^r) \\ &\quad + M_{J\tilde{J}} x_{\tilde{J}}^r + q_J = 0, \end{aligned}$$

or equivalently,

$$0 = Fx_J^{r+1} + Gx_J^r + H(x_{\tilde{J}}^{r+1} - x_{\tilde{J}}^r) + M_{J\tilde{J}} x_{\tilde{J}}^r + q_J,$$

for suitably defined matrices F and H, with $G = M_{JJ} - F$. Fix any $x^* \in X^*$. By subtracting the identity $0 = M_J x^* + q_J$ (cf. $J \subseteq I^*$ and Lemma 1) from the above equation and rearranging terms, we obtain

$$x_J^{r+1} - x_J^* = (I - F^{-1}M_{JJ})(x_J^r - x_J^*) - F^{-1}M_{J\bar{J}}(x_{\bar{J}}^r - x_{\bar{J}}^*) - F^{-1}H(x_{\bar{J}}^{r+1} - x_{\bar{J}}^r),$$

for $r = t, \cdots, s-1$. Now the matrix difference $F - G$ can be seen to have the form $L + E - L^T$, where L is a certain strictly (block) lower triangular part of M_{JJ} and E is a block diagonal matrix whose jth diagonal block is $B_{J_jJ_j} - C_{J_jJ_j}$. Therefore $\langle z, (F-G)z\rangle = \langle z, Ez\rangle > 0$ for all $z \neq 0$, where the strict inequality follows from the positive definite property of the $B_{J_jJ_j} - C_{J_jJ_j}$'s. This shows that (F, G) is a regular splitting of M_{JJ}. The rest of the proof then proceeds as in the proof of Lemma 8. □

Remark 6. We can also introduce under/overrelaxation in the GS-MS algorithm. More precisely, for each $j \in \{1, \cdots, m\}$, let $(B_{I_jI_j}, C_{I_jI_j})$ be a splitting of $M_{I_jI_j}$ and $\bar{\omega}_j$ be a scalar in $(0, 1]$ satisfying

$$B_{I_jI_j} - C_{I_jI_j} + (1 - \bar{\omega}_j)M_{I_jI_j} \quad \text{is positive definite.}$$

We define $\mathcal{A}_j$ as in (4.5) (but with the above splitting) and let $\mathcal{R}_j : X \mapsto X$ be the *underrelaxation* mapping corresponding to $\mathcal{A}_j$:

$$\mathcal{R}_j(x) = (1 - \bar{\omega}_j)x + \bar{\omega}_j\mathcal{A}_j(x).$$

Then the underrelaxed GS-MS algorithm comprises applications of the mappings $\mathcal{R}_1, \cdots, \mathcal{R}_m$ in a cyclical manner:

$$x^{r+1} = (\mathcal{R}_m \circ \cdots \circ \mathcal{R}_2 \circ \mathcal{R}_1)(x^r), \qquad r = 0, 1, \cdots.$$

In the special case where $\bar{\omega}_1 = \cdots = \bar{\omega}_m = 1$, the above algorithm reduces to the GS-MS algorithm. Furthermore, we can introduce an overrelaxation mechanism at the end of each iteration,

$$x^{r+1} = (1 - \omega^r)x^r + \omega^r(\mathcal{R}_m \circ \cdots \circ \mathcal{R}_2 \circ \mathcal{R}_1)(x^r), \qquad r = 0, 1, \cdots,$$

where each ω^r is chosen such that $x^{r+1} \in X$ and $\underline{\omega} < \omega^r \leqq \bar{\omega}$. The relaxation parameters $\underline{\omega}$ and $\bar{\omega}$ are chosen such that $0 < \underline{\omega} \leqq \min\{1, \bar{\omega}\}$ and $K + (1 - \bar{\omega})M$ is positive definite, where K is the $n \times n$ block diagonal matrix whose diagonal blocks comprise the positive definite matrices $(B_{I_jI_j} - C_{I_jI_j} + (1 - \bar{\omega}_j)M_{I_jI_j})/\bar{\omega}_j$, $j = 1, \cdots, m$. We can also introduce a positive diagonal matrix in the definition of $\mathcal{A}_j$ as in (4.4). In the special case where $m = 1$ and $\bar{\omega}_1 = 1$, this latter algorithm reduces to the algorithm (4.1)-(4.4). Convergence of the above algorithms can be shown by modifying the proof of Theorems 2 and 3.

We can alternatively extend the GS-MS algorithm to allow under/overrelaxation (during the updating of each subset of coordinates), nondisjoint subsets I_j, and noncyclic order of relaxation. This leads to the following SOR type algorithm, which we call the SOR-MS algorithm. Let $I_1, \cdots, I_m$ be a finite collection of nonempty (not necessarily disjoint) subsets of $\{1, \cdots, n\}$ whose union equals $\{1, \cdots, n\}$. For each $j = 1, \cdots, m$, we choose a splitting $(B_{I_jI_j}, C_{I_jI_j})$ of $M_{I_jI_j}$ and a $\bar{\omega}_j > 0$ satisfying

$$(4.8) \qquad B_{I_jI_j} - C_{I_jI_j} + (1 - \bar{\omega}_j)M_{I_jI_j} \quad \text{is positive definite.}$$

We also choose a second relaxation parameter $\underline{\omega}_j$ satisfying

$$(4.9) \qquad 0 < \underline{\omega}_j \leqq \min\{1, \bar{\omega}_j\},$$

and define $\mathcal{P}_j : X \mapsto X$ to be the point-to-set mapping

$$(4.10) \qquad \mathcal{P}_j(x) = \{z \mid z = (1 - \omega)x + \omega\mathcal{A}_j(x),\ z \in X, \text{ for some } \underline{\omega}_j \leqq \omega \leqq \bar{\omega}_j\},$$

where $\mathcal{A}_j : X \mapsto X$ is the point-to-point mapping given by (4.5). The SOR-MS algorithm generates a sequence of iterates by successively applying the mappings $\mathcal{P}_1, \cdots, \mathcal{P}_m$ (but not necessarily in any fixed order):

SOR-MS ALGORITHM. Choose an $x^0 \in X$. Generate a sequence of vectors $x^0, x^1, \cdots$ in X by the formula

$$x^{r+1} \in \mathcal{P}_{j^r}(x^r), \qquad r = 0, 1, \cdots, \tag{4.11}$$

where $j^0, j^1, \cdots$ is some sequence of indices in $\{1, \cdots, m\}$.

We will impose the following rule on the order of coordinate relaxation (see, e.g., [SaS73], [HeL78]):

ALMOST CYCLIC RULE. There exists an integer $\bar{r}$ such that $\{1, \cdots, m\} \subseteq \{j^{r+1}, j^{r+2}, \cdots, j^{r+\bar{r}}\}$ for all r.

The SOR-MS algorithm can be seen to contain most of the earlier algorithms as special cases. For example, if $m = 1$, then it reduces to the algorithm (4.1)-(4.4). If the I_j's are disjoint, $\omega_j = \bar{\omega}_j = 1$ for all j, and $\{j^0, j^1, \cdots\} = \{1, \cdots, m, 1, \cdots, m, \cdots\}$, then it reduces to the GS-MS algorithm. It also contains other methods as special cases. For example, if the $M_{I_jI_j}$'s are positive definite and we choose $B_{I_jI_j} = M_{I_jI_j}$ for all j, then (4.8) is equivalent to $\bar{\omega}_j < 2$ and the SOR-MS algorithm reduces to a block SOR method considered in [Tse88, § 6.2]. Furthermore, if the I_j's are disjoint and $\{j^0, j^1, \cdots\} = \{1, \cdots, m, 1, \cdots, m, \cdots\}$, then it reduces to the block SOR methods considered in [CGS78], [CoG78]; and if $m = n$ and $I_j = \{j\}$ for all j, then it reduces to the point SOR methods of Herman and Lent [HeL78] and of Lent and Censor [LeC80]. For a final example, if $I_j = \{1, \cdots, n\}$ for all j, then it reduces to a matrix splitting algorithm that alternates amongst m matrix splittings.

We have not been able to show that the SOR-MS algorithm is convergent in the sense of Theorems 1-3. (The difficulty lies in the proof of Lemma 8, which no longer goes through when the index subsets $I_1, \cdots, I_m$ overlap or when different *and* time varying relaxation parameters are placed on different coordinates.) However, by combining the second half of the proof of Theorem 2 with the first half of the proof of Theorem 3, we can show that it is convergent in the weaker sense of Lemma 7.

THEOREM 4. *Let* $x^0, x^1, \cdots$ *denote the iterates generated by the* SOR-MS *algorithm* (4.5), (4.8)-(4.11) *under the almost cyclic rule. Then* $Mx^r + q \to d^*$ *and* $\phi(x^r) \to 0$. *Moreover,* $f(x^r)$ *tends to the optimal value of* (P) *and every limit point of* $\{x^r\}$ *is a solution of* (P).

Although the above result is not the strongest one possible, it nonetheless improves upon those existing. For example, it shows, for the first time, that the algorithms considered in [Tse88, § 6.2], [HeL78], [LeC80], [CGS78], and [CoG78] generate iterates that come arbitrarily close to the solution set X^*.

5. Application to known methods. In this section we apply the results developed in §§ 3 and 4 to a number of well-known methods and show, for the first time, that these methods are convergent without making any additional assumption on the problem. We also extend some of these methods to incorporate overrelaxation.

Example 1 (point SOR method). Suppose that M has positive diagonal entries. Consider the following well-known point SOR method [Hil57], [Cry71], [Man84] for solving (P):

$$x_i^{r+1} = \left[x_i^r - \frac{\alpha}{M_{ii}} \left(\sum_{j<i} M_{ij} x_j^{r+1} + \sum_{j \geq i} M_{ij} x_j^r + q_i \right) \right]^+, \qquad i = 1, \cdots, n,$$

where α is a relaxation parameter in $(0, 2)$ and $[\,\cdot\,]^+$ denotes the orthogonal projection onto the interval $[0, \infty)$. (This method can be viewed alternatively as a (cyclic) coordinate descent method with inexact line search [Tse88, § 6.2].) It is easily seen that this method is a special case of the algorithm (4.1)-(4.4) with $\underline{\omega} = \bar{\omega} = 1$ and the following choices of (B, C) and D:

$$B = \alpha^{-1}E + L, \quad C = (1 - \alpha^{-1})E + L^T, \quad D = \alpha E^{-1},$$

where E and L are, respectively, the diagonal and the strictly lower triangular part of M. Since $B - C = (2\alpha^{-1} - 1)E + L - L^T$, which is positive definite for all $\alpha \in (0, 2)$, it follows from Theorem 2 that this method is convergent. This improves upon existing results (e.g., [Cry71], [Man84], [LiP87]), which require for convergence either M be *strictly copositive* or that a certain Slater condition hold (all of which lead to the compactness of X^*).

Example 2 (gradient projection algorithms). Consider the well-known gradient projection algorithm [Gol64], [LeP65] (also see [Ber82], [BeT89], [Che84], [Lue73]) applied to solve (P),

$$x^{r+1} = [x^r - \alpha(Mx^r + q)]^+,$$

where α is a positive stepsize. It is easily seen that this is a special case of the algorithm (4.1)-(4.4) with $\underline{\omega} = \bar{\omega} = 1$ and the following choices of (B, C) and D:

$$B = \frac{1}{\alpha} I, \quad C = M - \frac{1}{\alpha} I, \quad D = \alpha I.$$

In this case $B - C$ can be seen to be positive definite for all $\alpha < 2/\|M\|$. Hence by Theorem 1, the algorithm is convergent for all $\alpha \in (0, 2/\|M\|)$. (This result was first established by Cheng [Che84] for the more general problem of minimizing a pseudoconvex differentiable function over a closed convex set.) Aganagic [Aga78] proposed a modification of the above algorithm by adding a relaxation parameter $\omega \in (0, 1]$:

$$x^{r+1} = (1 - \omega)x^r + \omega[x^r - \alpha(Mx^r + q)]^+.$$

This algorithm is also a special case of the algorithm (4.1)-(4.4) with $\underline{\omega} = \bar{\omega} = \omega$ and with (B, C) and D given as above. Hence, by Theorem 2, this algorithm is also convergent for all $\alpha \in (0, 2/\|M\|)$. (This improves on the result of Aganagic which requires M to be positive definite for convergence. Furthermore, from Theorem 2 we see that overrelaxation (i.e., $\omega > 1$) is also permissible, as long as $\alpha\omega \in (0, 2/\|M\|)$.)

Example 3 (Mangasarian's algorithm). Consider the following iterative algorithm proposed by Mangasarian [Man77] (also see [Man84], [MaD87] for applications)

$$x^{r+1} = (1 - \omega)x^r + \omega[x^r - \alpha E(Mx^r + q + K(x^{r+1} - x^r))]^+,$$

where $\omega \in (0, 1]$, E is an $n \times n$ positive diagonal matrix, K is an $n \times n$ matrix, and α is a positive scalar. It can be seen that the above algorithm is a special case of the algorithm (4.1)-(4.4) with $\underline{\omega} = \bar{\omega} = \omega$ and the following choices of (B, C) and D:

$$B = (\alpha E)^{-1} + \omega K, \quad C = M - (\alpha E)^{-1} - \omega K, \quad D = \alpha E.$$

Since $B - C + (1 - \omega)M = 2(\alpha E)^{-1} + 2\omega K - \omega M$, it follows from Theorem 2 that the above algorithm is well defined and convergent if $2(\alpha E)^{-1} + 2\omega K - \omega M$ is positive definite, which is exactly the condition given by Mangasarian [Man77, (6)]. (Mangasarian proved that the algorithm itself is well defined for all choices of the matrix K satisfying either his assumption [Man77, (6)], or the assumption that K is strictly lower triangular, which was the case of principal concern in [Man77]. He also showed that each limit point of the iterates generated by the algorithm is a solution, but the question of whether such a limit point exists was left open.)

Example 4 (block SOR method). Consider the following block SOR method of Cottle, Golub, and Sacher [CGS78] and of Cottle and Goheen [CoG78] (also see [CoP82]). Partition the index set $\{1, \cdots, n\}$ into m nonempty, mutually disjoint subsets $I_1, \cdots, I_m$ and assume that $M_{I_jI_j}$ is positive definite for all j. Choose a relaxation parameter $\bar{\omega} \in (0, 2)$. Then, for any given $x^0 \in X$, the method generates a sequence of iterates $\{x^0, x^1, \cdots\}$ whereby, given x^r, a new iterate x^{r+1} is generated as follows.

Let $z^0 = x^r$. For $j = 1, \cdots, m$, compute $\hat{z}^j$ to be the (unique) solution to the following system of nonlinear equations ($[\cdot]^+$ denotes the orthogonal projection onto the interval $[0, \infty)$)

$$z_i = [z_i - (M_i z + q_i)]^+ \quad \forall i \in I_j,$$
$$z_i = z_i^{j-1} \quad \forall i \notin I_j,$$

and let $z^j = (1-\omega)z^{j-1} + \omega \hat{z}^j$, where ω is the largest scalar in $(0, \bar{\omega}]$ such that $z^j \in X$. Then set $x^{r+1} = z^m$. (This method essentially replaces the strictly lower triangular (diagonal) part of M in the point SOR method by strictly lower triangular (diagonal) blocks.) In the case where $\bar{\omega} = 1$, this method can be seen to be a special case of the GS-MS algorithm (4.5)-(4.6) with

$$B_{I_jI_j} = M_{I_jI_j}, \qquad C_{I_jI_j} = 0,$$

so that by Theorem 3 it is convergent. If $0 < \bar{\omega} \leqq 1$, then by Remark 6 it is also convergent. (This improves upon the results of [CGS78] and [CoG78] which require M to be positive definite for convergence. It also obviates the need for the projection step employed in [CoP82] to ensure the existence of a limit point.) In the case where $1 < \bar{\omega} < 2$, however, the convergence of this method remains unresolved. It is known to be convergent only in the weak sense of Theorem 4.

6. Discussions and extensions. In this paper, we have established the iterate convergence of matrix splitting algorithms for solving the symmetric monotone linear complementarity problem when the splitting is regular. Our result improves on the earlier convergence results in that it does not make any assumption on the boundedness of the optimal solution set. Our proof makes essential use of linearity of the problem and certain contractive properties of the iteration matrices. In particular, these contractive properties enable us to carefully estimate the behaviors of the algorithm near the boundary of the feasible set.

There are several directions in which our results can be improved. For example, we may consider solving each of the subproblems (1.5) *inexactly*. Specifically, consider the sequence of iterates $\{x^1, x^2, \cdots,\}$ generated by

$$x^{r+1} = \mathcal{A}_B(x^r + e^r),$$

where e^r denotes the "error" vector at the rth iteration. Recently, Mangasarian [Man90] established the convergence of $\{x^r\}$ under (essentially) the assumption that B is symmetric and that the error vectors $\{e^r\}$ satisfy

$$\sum_{r=0}^{\infty} \|e^r\| < \infty, \tag{6.1}$$

$$\sum_{r=0}^{\infty} \|e^r\| \, \|x^{r+1} - x^r\| < \infty. \tag{6.2}$$

We remark that our proof (Lemmas 8-11) can be modified to establish the convergence of $\{x^r\}$ for any regular splitting (B, C) (not necessarily symmetric) under the same set of assumptions (6.1) and (6.2). This is because each error vector e^r contributes additively an $O(\|e^r\|)$ perturbation to all future iterates (cf. Lemma 4(b)), so the perturbation contributed by all of the error vectors after the rth iteration is only $O(\sum_{i=r}^{\infty} \|e^i\|)$, which, by (6.1), tends to zero as $r \to \infty$. (Condition (6.2) is needed to ensure that Lemma 6 still holds (see [Man90]).)

Another interesting extension of the results obtained here is to analyze the *rate* of convergence of the matrix splitting algorithms. Such an analysis has been obtained by Luo and Tseng [LuT89] for the special case of the coordinate descent method (but for the more general problem of minimizing, over a box, a function which is the composition of an affine mapping with a strictly convex essentially smooth function) and only recently we have been able to extend their analysis to matrix splitting algorithms.

Acknowledgments. We express our deep gratitude to the referees for providing many valuable comments and for bringing the work of Keller [Kel65] to our attention. We also thank Professor O. L. Mangasarian for his helpful suggestions.

REFERENCES

[AdG75] I. ADLER AND D. GALE, *On the solution of the positive semi-definite complementarity problem*, Tech. Report 75-12, Operations Research Center, University of California, Berkeley, CA, 1975.

[Aga78] M. AGANAGIC, *Iterative methods for linear complementarity problems*, Tech. Report SOL 78-10, Systems Optimization Laboratory, Department of Operations Research, Stanford University, Stanford, CA, 1978.

[BeP79] A. BERMAN AND R. J. PLEMMONS, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[Ber82] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[BHT87] D. P. BERTSEKAS, P. A. HOSEIN, AND P. TSENG, *Relaxation methods for network flow problems with convex arc costs*, SIAM J. Control Optim., 25 (1987), pp. 1219-1243.

[BeT89] D. P. BERTSEKAS AND J. N. TSITSIKLIS, *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[CeG73] J. CEA AND R. GLOWINSKI, *Sur des méthodes d'optimisation par relaxation*, Revue Française d'Automatique, Informatique et Recherche Opérationelle, R-3 (1973), pp. 5-31.

[Che84] Y. C. CHENG, *On the gradient-projection method for solving the nonsymmetric linear complementarity problem*, J. Appl. Math. Optim., 43 (1984), pp. 527-540.

[CoG78] R. W. COTTLE AND M. S. GOHEEN, *A special class of large quadratic programs*, in Nonlinear Programming 3, O. L. Mangasarian, R. R. Meyer, and S. M. Robinson, eds., Academic Press, New York, 1978, pp. 361-390.

[CGS78] R. W. COTTLE, G. H. GOLUB, AND R. S. SACHER, *On the solution of large, structured linear complementarity problems: the block partitioned case*, J. Appl. Math. Optim., 4 (1978), pp. 347-363.

[CoP82] R. W. COTTLE AND J.-S. PANG, *On the convergence of a block successive over-relaxation method for a class of linear complementarity problems*, Math. Programming Stud., 17 (1982), pp. 126-138.

[Cry71] C. W. CRYER, *The solution of a quadratic programming problem using systematic overrelaxation*, SIAM J. Control, 9 (1971).

[DeT84] R. S. DEMBO AND U. TULOWITZKI, *On the minimization of quadratic functions subject to box constraints*, School of Organization and Management and Department of Computer Science, Yale University, New Haven, CT, 1983 (revised 1984).

[D'Es59] D. A. D'ESOPO, *A convex programming procedure*, Naval Res. Logist. Quart., 6 (1959), pp. 33-42.

[Eav71] B. C. EAVES, *On quadratic programming*, Management Sci., 17 (1971), pp. 698-711.

[FrW57] M. FRANK AND P. WOLFE, *An algorithm for quadratic programming*, Naval Res. Logist. Quart., 3 (1957), pp. 95-110.

[Gol64] A. A. GOLDSTEIN, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709-710.

[HeL78] G. T. HERMAN AND A. LENT, *A family of iterative quadratic optimization algorithms for pairs of inequalities, with application in diagnostic radiology*, Math. Programming Stud., 9 (1978), pp. 15-29.

[Hil57] C. HILDRETH, *A quadratic programming procedure*, Naval Res. Logist. Quart., 4 (1957), pp. 79-85; Erratum, Naval Res. Logist. Quart., 4 (1957), p. 361.

[Kel65] H. B. KELLER, *On the solution of singular and semi-definite linear systems by iteration*, SIAM J. Numer. Anal., 2 (1965), pp. 281-290.

[KiS80] D. KINDERLEHRER AND G. STAMPACCHIA, *An Introduction to Variational Inequalities and Applications*, Academic Press, New York, 1980.

[LeC80] A. LENT AND Y. CENSOR, *Extensions of Hildreth's row—action method for quadratic programming*, SIAM J. Control Optim., 18 (1980), pp. 444-454.

[LeP65] E. S. LEVITIN AND B. T. POLJAK, *Constrained minimization methods*, Zh. Vychisl. Mat. i Mat. Fiz., 6 (1965), pp. 787-823. (In Russian.) USSR Comput. Math. and Math. Phys., 6 (1965), pp. 1-50. (In English.)

[LiP87] Y. Y. LIN AND J.-S. PANG, *Iterative methods for large convex quadratic programs: a survey*, SIAM J. Control Optim., 25 (1987), pp. 383-411.

[Lue73] D. G. LUENBERGER, *Linear and Nonlinear Programming*, Addison-Wesley, Reading, MA, 1973.

[LuT89] Z.-Q. LUO AND P. TSENG, *On the convergence of the coordinate descent method for convex differentiable minimization*, LIDS Report P-1924, Massachusetts Institute of Technology, Cambridge, MA, December 1989; J. Optim. Theory Appl., to appear.

[Man77] O. L. MANGASARIAN, *Solution of symmetric linear complementarity problems by iterative methods*, J. Optim. Theory Appl., 22 (1977), pp. 465-485.

[Man84] ———, *Sparsity-preserving* SOR *algorithms for separable quadratic and linear programming*, Comput. Oper. Res., 11 (1984), pp. 105-112.

[Man88] ———, *A simple characterization of solution sets of convex programs*, Oper. Res. Lett., 7 (1988), pp. 21-26.

[Man90] ———, *Convergence of iterates of an inexact matrix splitting algorithm for the symmetric monotone linear complementarity problem*, Computer Sciences Tech. Report #917, University of Wisconsin, Madison, WI, March 1990.

[MaD87] O. L. MANGASARIAN AND R. DE LEONE, *Parallel successive overrelaxation methods for symmetric linear complementarity problems and linear programs*, J. Optim. Theory Appl., 54 (1987), pp. 437-446.

[OrR70] J. M. ORTEGA AND W. C. RHEINBOLDT, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.

[Pan82] ———, *On the convergence of a basic iterative method for the implicit complementarity problem*, J. Optim. Theory Appl., 37 (1982), pp. 149-162.

[Pan84] ———, *Necessary and sufficient conditions for the convergence of iterative methods for the linear complementarity problem*, J. Optim. Theory Appl., 42 (1984), pp. 1-17.

[Pan86] ———, *More results on the convergence of iterative methods for the symmetric linear complementarity problem*, J. Optim. Theory Appl., 49 (1986), pp. 107-134.

[Rob81] S. M. ROBINSON, *Some continuity properties of polyhedral multifunctions*, Math. Prog. Stud., 14 (1981), pp. 206-214.

[SaS73] R. W. H. SARGENT AND D. J. SEBASTIAN, *On the convergence of sequential minimization algorithms*, J. Optim. Theory Appl., 12 (1973), pp. 567-575.

[Tse88] P. TSENG, *Dual ascent methods for problems with strictly convex costs and linear constraints: a unified approach*, SIAM J. Control Optim., 28 (1990), pp. 214-242.

[Tse89] ———, *Further applications of a splitting algorithm to decomposition in variational inequalities and convex programming*, Math. Programming B, 48 (1990), pp. 249-263.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 5, pp, 1061-1090, September 1991

RECURSIVE IDENTIFICATION AND ADAPTIVE PREDICTION IN LINEAR STOCHASTIC SYSTEMS*

TZE LEUNG LAI† AND ZHILIANG YING‡

Abstract. By making use of extended stochastic Lyapunov functions and martingale limit theorems, established herein are certain basic properties of adaptive d-step ahead predictors associated with the extended least squares, stochastic gradient (without interlacing), and monitored recursive maximum likelihood algorithms for recursive identification of an ARMAX system. Both the direct (or implicit) and indirect (or explicit) approaches to adaptive prediction are considered within a unified framework involving stochastic regression models. Applications to adaptive control of ARMAX systems are also discussed.

Key words. adaptive prediction, global convergence, stochastic gradient algorithm, AML, recursive MLE, stochastic adaptive control, certainty equivalence, asymptotic efficiency

AMS(MOS) subject classifications. 93E12, 93C40, 60G42

1. Introduction and background. Consider the ARMAX system (autoregressive moving average system with exogenous inputs) defined by the linear stochastic difference equation

$$A(q^{-1})y_n = q^{-\Delta}B(q^{-1})u_n + C(q^{-1})\varepsilon_n, \tag{1.1}$$

where $\{y_n\}$, $\{u_n\}$, and $\{\varepsilon_n\}$ denote the output, input, and disturbance sequences, respectively, $\Delta \geqq 1$ represents the delay, and

$$\begin{aligned} &A(q^{-1}) = 1 + a_1 a^{-1} + \cdots + a_p q^{-p}, \qquad B(q^{-1} = b_1 + \cdots + b_k q^{-(k-1)}, \\ &C(q^{-1}) = 1 + c_1 q^{-1} + \cdots + c_h q^{-h} \end{aligned} \tag{1.2}$$

are scalar polynomials in the backward shift operator q^{-1}. Throughout the sequel we will assume that the sequence $\{\varepsilon_n\}$ is a martingale difference sequence with respect to an increasing sequence of σ-fields $\mathscr{F}_n$ such that

$$\sup_n E(|\varepsilon_n|^\alpha \mid \mathscr{F}_{n-1}) < \infty \quad \text{a.s. for some } \alpha > 2. \tag{1.3}$$

Moreover, the input u_t at stage t is assumed to be $\mathscr{F}_t$-measurable (i.e., involving only the current and past observations $y_t, y_{t-1}, u_{t-1}, \cdots$, but no future observations). Letting $x_0 = (y_0, \cdots, y_{1-p}, u_0, \cdots, u_{2-\Delta-k}, \varepsilon_0, \cdots, \varepsilon_{1-h})$ denote the "initial condition" of (1.1), it is also assumed that x_0 is $\mathscr{F}_0$-measurable.

Let $1 \leqq d \leqq \Delta$. When the system parameters $a_1, \cdots, a_p, b_1, \cdots, b_k, c_1, \cdots, c_h$ and the initial condition x_0 are known, the minimum variance d-step ahead predictor $\tilde{y}_{n+d} \triangleq E(y_{n+d} \mid \mathscr{F}_n)$ of the output y_{n+d} can be determined recursively by the Åström predictor identity (1.6) below (cf. [1]-[4]). By the division algorithm, there exist polynomials $F(z) = 1 + f_1 z + \cdots + f_{d-1} z^{d-1}$ and $G(z) = g_1 + \cdots + g_{p(d)} z^{p(d)-1}$ with $p(d) = p \vee (h - d + 1)$ such that

$$C(z) = F(z)A(z) + z^d G(z), \tag{1.4}$$

* Received by the editors November 27, 1989; accepted for publication (in revised form) November 26, 1990. This work was supported by the National Science Foundation, the National Security Agency, and the Air Force Office of Scientific Research.

† Department of Statistics, Stanford University, Stanford, California 94305.

‡ Department of Statistics, University of Illinois, Champaign, Illinois 61820.

and therefore (1.1) can be rewritten in the form

$$C(q^{-1})\{y_{n+d}-F(q^{-1})\varepsilon_{n+d}\}=G(q^{-1})y_n+q^{-(\Delta-d)}F(q^{-1})B(q^{-1})u_n. \tag{1.5}$$

This implies that the minimum variance predictor $\tilde{y}_{n+d}$ is given recursively by

$$C(q^{-1})\tilde{y}_{n+d}=G(q^{-1})y_n+q^{-(\Delta-d)}F(q^{-1})B(q^{-1})u_n. \tag{1.6}$$

The prediction error of the predictor $\tilde{y}_{n+d}$ is

$$\eta_{n+d}\overset{\Delta}{=}y_{n+d}-\tilde{y}_{n+d}=F(q^{-1})\varepsilon_{n+d}. \tag{1.7}$$

In practice, the system parameters and initial condition are usually unknown, and we must "adapt" the optimal predictor (1.6) by substituting the unknown entities in (1.6) by their estimates. The so-called *explicit* (or *indirect*) approach of adaptive prediction is to first estimate the parameters $a_1,\cdots,a_p, b_1,\cdots,b_k, c_1,\cdots,c_h$ of the explicit dynamical system (1.1) and then to substitute these parameter values that appear in the polynomials $C(q^{-1})$, $B(q^{-1})$, $F(q^{-1})$, and $G(q^{-1})$ of (1.6) by their estimates. In contrast, the *implicit* (or *direct*) approach of adaptive prediction is to first develop recursive estimates θ_n of the parameter vector

$$\begin{gathered}\theta=(g_1,\cdots,g_{p(d)},b_1,(fb)_2,\cdots,(fb)_{k+d-1},-c_1,\cdots,-c_h)', \quad \text{where}\\ \sum_{i=1}^{k+d-1}(fb)_i z^{i-1}=F(z)B(z), \quad \text{so that } (fb)_1=b_1,\end{gathered} \tag{1.8}$$

of the system's implicit representation that combines (1.5) and (1.7) into the form

$$\begin{gathered}y_{n+d}=\theta'\psi_n+\eta_{n+d}, \quad \text{where}\\ \psi_n=(y_n,\cdots,y_{n-p(d)+1},u_{n-\Delta+d},\cdots,u_{n-k-\Delta+2},\tilde{y}_{n+d-1},\cdots,\tilde{y}_{n+d-h})',\end{gathered} \tag{1.9}$$

noting that $\tilde{y}_{n+d}=\theta'\psi_n$ by (1.6). Letting $\hat{y}_{n+d}$ denote the adaptive predictor of y_{n+d}, this implicit approach generates $\hat{y}_{n+d}$ recursively by

$$\begin{gathered}\hat{y}_{n+d}=\theta_n'\phi_n, \quad \text{where}\\ \phi_n=(y_n,\cdots,y_{n-p(d)+1},u_{n-\Delta+d},\cdots,u_{n-k-\Delta+2},\hat{y}_{n+d-1},\cdots,\hat{y}_{n+d-h})'.\end{gathered} \tag{1.10}$$

For the explicit approach, there is a large literature on recursive estimation of the parameter vector

$$\Theta=(-a_1,\cdots,-a_p,b_1,\cdots,b_k,c_1,\cdots,c_h)' \tag{1.11}$$

of the dynamical system (1.1), which can be written in the regression form

$$\begin{gathered}y_n=\Theta'\Psi_{n-1}+\varepsilon_n, \quad \text{where}\\ \Psi_t=(y_t,\cdots,y_{t-p+1},u_{t-\Delta+1},\cdots,u_{t-\Delta-k+2},\varepsilon_t,\cdots,\varepsilon_{t-h+1})'.\end{gathered} \tag{1.12}$$

The recent monographs [2]-[4] provide excellent unified overviews of various recursive estimation algorithms in the literature. In particular, an important and widely studied problem concerning, these recursive estimators is under what conditions they converge. In the seminal papers [5], [6], Ljung developed the ODE (ordinary differential equation) method for the convergence analysis of recursive estimators Θ_n of Θ. Under certain a priori boundedness and recurrence assumptions on Θ_n, this method introduces a space-time renormalization into the recursion for $\Theta_n-\Theta$ to obtain a nonrandom ODE as a limit point and studies the limiting behavior of Θ_n via the stability properties of the associated ODE. Such stability analysis is often conveniently carried out by making use of a Lyapunov function. Instead of working with a Lyapunov function associated with the limiting ODE, an obvious alternative is to develop an analogue for the original recursions defining Θ_n. This is the idea behind the "stochastic Lyapunov function" approach introduced by Moore and Ledwich [7] and Solo [8]. A basic ingredient of this approach is to use the underlying system dynamics to develop recursive inequalities for a suitably chosen nonnegative random function of Θ_n and to normalize and transform this function into a nonnegative almost supermartingale (stochastic Lyapunov function) to which the martingale convergence theorem can be applied. In particular, for the AML algorithm

$$\Theta_n=\Theta_{n-1}+P_{n-1}\Phi_{n-1}(y_n-\Theta_{n-1}'\Phi_{n-1}),\qquad P_n^{-1}=P_{n-1}^{-1}+\Phi_n\Phi_n', \tag{1.13a}$$

$$\Phi_n=(y_n,\cdots,y_{n-p+1},u_{n-\Delta+1},\cdots,u_{n-\Delta-k+2},\hat{\varepsilon}_n,\cdots,\hat{\varepsilon}_{n-h+1})', \tag{1.13b}$$

$$\hat{\varepsilon}_n=y_n-\Theta_n'\Phi_{n-1}, \tag{1.13c}$$

Solo [8] used this approach to prove the strong consistency of Θ_n under both the "persistent excitation" condition

$$n^{-1}\sum_{i=1}^{n}\Psi_i\Psi_i' \quad \text{converges a.s. to a positive-definite matrix} \tag{1.14}$$

and the "positive real" condition

$$\min_{|z|=1}\operatorname{Re}\left(\frac{1}{C(z)}-\frac{1}{2}\right)>0. \tag{1.15}$$

For the AML algorithm, by using martingale limit theorems (not restricted to convergence) to analyze directly Solo's recursive inequalities for the quadratic form

$$Q_n\triangleq(\Theta_n-\Theta)'P_{n-1}^{-1}(\Theta_n-\Theta), \tag{1.16}$$

instead of following Solo [8] to transform Q_n into a nonnegative almost supermartingale that converges almost surely by the martingale convergence theorem, Lai and Wei [9] established the strong consistency of Θ_n under (1.15) and the much weaker excitation condition

$$\lambda_{\min}\left(\sum_{i=1}^{n}\Psi_i\Psi_i'\right)\to\infty \quad\text{and}\quad \log\lambda_{\max}\left(\sum_{i=1}^{n}\Psi_i\Psi_i'\right)=o\left(\lambda_{\min}\left(\sum_{i=1}^{n}\Psi_i\Psi_i'\right)\right)\quad\text{a.s.} \tag{1.17}$$

Here and in the sequel we use $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to denote the maximum and minimum eigenvalues of a symmetric matrix A. Because of the Lyapunov-type recursive inequalities satisfied by Q_n which, however, need not be convergent, we will call such functions "extended stochastic Lyapunov functions" as in [10], where it is shown that stronger results can often be obtained by applying martingale theory directly to such functions without transforming them into (convergent) stochastic Lyapunov functions.

In the explicit approach to adaptive prediction, for a recursive algorithm Θ_n estimating the unknown parameter vector Θ defined in (1.11), we first use Θ_n at stage n to estimate the unknown coefficients of the polynomials $C(q^{-1})$, $B(q^{-1})$, $F(q^{-1})$, and $G(q^{-1})$ in (1.6), leading to the estimated polynomials $\hat{C}_n(q^{-1})$, $\hat{B}_n(q^{-1})$, $\hat{F}_n(q^{-1})$, and $\hat{G}_n(q^{-1})$ at stage n, and then define the predictor $\hat{y}_{n+d}$ of y_{n+d} by the recursive relation

$$\hat{C}_n(q^{-1})\hat{y}_{n+d}=\hat{G}_n(q^{-1})y_n+q^{-(\Delta-d)}\hat{F}_n(q^{-1})\hat{B}_n(q^{-1})u_n, \tag{1.18}$$

noting that the coefficients of $F(q^{-1})$ and $G(q^{-1})$ are polynomial functions of the components of Θ by (1.4). Hence, if Θ_n converges almost surely to Θ and $C(z)$ is stable (i.e., all its zeros lie outside the unit circle), then it follows from (1.16) and (1.18) that

$$\sum_{i=1}^{n} (\tilde{y}_{i+d} - \hat{y}_{i+d})^2 = o\left(\sum_{i=1}^{n} (y_i^2 + u_i^2)\right) \quad \text{a.s.} \tag{1.19}$$

(cf. Lemma 5 of § 2). In most applications, we typically have sample mean square boundedness for the input-output data, i.e.,

$$\limsup_{n\to\infty} n^{-1} \sum_{i=1}^{n} (y_i^2 + u_i^2) < \infty \quad \text{a.s.,} \tag{1.20}$$

in which case (1.19) implies that

$$n^{-1} \sum_{i=1}^{n} (\tilde{y}_{i+d} - \hat{y}_{i+d})^2 \to 0 \quad \text{a.s.} \tag{1.21}$$

A sequence of d-step ahead predictors $\{\hat{y}_{n+d}\}$ is said to be "globally convergent" if (1.21) holds (cf. [11]). We have pointed out above that if a consistent estimator Θ_n of the parameter vector Θ of the explicit system (1.11) can be found and if $C(z)$ is stable and (1.20) holds, then globally convergent adaptive predictors can be constructed by the recursive relation (1.18). However, the requirement of consistency in parameter estimation is often not needed in the construction of globally convergent adaptive predictors, particularly if we use an implicit approach. Extending the AML algorithm to the implicit model (1.9), Sin, Goodwin, and Bitmead [12] constructed d-step ahead adaptive predictors based on an interlaced AML algorithm and showed that such predictors are globally convergent under assumption (1.20) and an assumption analogous to (1.15). As pointed out by Zhang [13], however, their proof uses Solo's [8] result (A6) whose proof contains a gap. In § 3, where Solo's result (A6) is shown to be incorrect without additional assumptions, we prove a stronger result than the Sin-Goodwin-Bitmead theorem under the additional assumption that $\limsup_{n\to\infty} \phi_n'(\sum_1^n \phi_i\phi_i')^{-1}\phi_n < 1$ almost surely, where the ϕ_n are the pseudoregression vectors in their algorithm.

The global convergence property (1.21) for adaptive predictors is of particular interest in the adaptive control problem of setting the input u_t at stage t so that the output $y_{t+\Delta}$ is as close as possible to some target value $y_{t+\Delta}^*$. When the system parameters and the initial condition x_0 are known and $b_1 \neq 0$, the minimum variance controller is to set u_t such that $\tilde{y}_{t+\Delta} = y_{t+\Delta}^*$. In ignorance of x_0 and the system parameters, it is therefore natural to set u_t such that $\hat{y}_{t+\Delta} = y_{t+\Delta}^*$, where $\hat{y}_{t+\Delta}$ is a globally convergent adaptive Δ-step ahead predictor of $y_{t+\Delta}$. Since $\tilde{y}_{t+\Delta} = y_{t+\Delta} - \eta_{t+\Delta}$, (1.21) implies the so-called "self-optimizing" property that

$$n^{-1} \sum_{i=\Delta+1}^{n} (y_i - y_i^* - \eta_i)^2 \to 0 \quad \text{a.s.,} \tag{1.22}$$

for the adaptive controller defined by $\hat{y}_{t+\Delta} = y_{t+\Delta}^*$.

In the case where unit delay $\Delta = 1$, Goodwin, Ramadge, and Caines [14] used this approach in conjunction with the stochastic gradient algorithm

$$\theta_n = \theta_{n-1} + (a/r_{n-1})\phi_{n-1}(y_n - \hat{y}_n), \qquad r_n = r_{n-1} + \|\phi_n\|^2, \tag{1.23}$$

where ϕ_t and $\hat{y}_t$ are defined in (1.10) (with $d = 1$) and $a > 0$, to establish the self-optimizing property (1.22) for the adaptive controller that chooses the input u_t so that

$$\hat{y}_{t+1} = y_{t+1}^*, \tag{1.24}$$

under the assumptions

$$\min_{|z|=1} \operatorname{Re}\left(C(z)-\frac{a}{2}\right)>0, \tag{1.25}$$

$$B(z) \text{ is stable and } b_1 \neq 0, \tag{1.26}$$

(1.27) $(x_0, \varepsilon_1, \cdots, \varepsilon_n)$ is absolutely continuous with respect to Lebesgue measure for every $n \geqq 1$.

Assumption (1.27) ensures that the component of θ_n estimating the component b_1 of θ is nonzero almost surely and therefore we can indeed define u_t by $\theta_t' \phi_t = y_{t+1}^*$, i.e., by (1.24). As shown in [11] for general delay Δ but still for $d=1$, (1.25) and (1.20) are sufficient conditions for the global convergence of adaptive one-step ahead predictors based on the stochastic gradient algorithm (1.23). Assumption (1.26) is needed to ensure that (1.20) holds for the adaptive controller (1.24).

For general delay Δ and $d=\Delta$, Goodwin, Sin, and Saluja [15] extended (1.23) to the form

$$\theta_n = \theta_{n-d} + (a/r_{n-d})\phi_{n-d}(y_n - \theta_{n-d}'\phi_{n-d}), \qquad r_n = r_{n-1} + \|\phi_n\|^2, \tag{1.28}$$

in which ϕ_n is given by (1.10). Under assumptions (1.25) and (1.20), they showed that the adaptive d-step head predictors $\hat{y}_{n+d} = \theta_n' \phi_n$ are globally convergent. By using the explicit instead of the implicit approach, Fuchs [16], [17] also constructed globally convergent adaptive d-step ahead predictors of the form (1.18), in which the coefficients of the polynomials $\hat{C}_n(q^{-1})$, $\hat{B}_n(q^{-1})$, $\hat{F}_n(q^{-1})$, and $\hat{G}_n(q^{-1})$ are determined from the stochastic gradient algorithm Θ_n estimating (1.11), defined by

$$\Theta_n = \Theta_{n-1} + (a/r_{n-1})\Phi_{n-1}\, e_n, \tag{1.29a}$$

$$e_n = y_n - \Theta_{n-1}'\Phi_{n-1}, \tag{1.29b}$$

$$\begin{aligned} &r_n = r_{n-1} + \|\Phi_n\|^2, \\ &\text{where } \Phi_n = (y_n, \cdots, y_{n-p+1}, u_{n-\Delta+1}, \cdots, u_{n-\Delta-k+2}, e_n, \cdots, e_{n-h+1})'. \end{aligned} \tag{1.29c}$$

Unlike the single recursion in (1.29), the algorithm (1.28) interlaces d recursions $\theta_{j+d(t+1)} = \theta_{j+dt} + (a/r_{j+dt})\phi_{j+dt}(y_{j+d(t+1)} - \theta_{j+dt}'\phi_{j+dt})$, $j=0, \cdots, d-1$. It has been an open problem concerning whether such multiple recursions are indeed necessary and not just dictated by the stochastic Lyapunov function method of convergence analysis (cf. [16, p. 219]). By using the extended stochastic Lyapunov function approach instead, we show in § 4 that interlacing is not needed to establish global convergence of adaptive d-step ahead predictors based on the stochastic gradient algorithm for the implicit model (1.9), giving a positive answer to this long-standing open problem.

Although the stochastic gradient algorithm (with a scalar gain a/r_t) leads to globally convergent adaptive predictors by either the explicit or implicit approach under assumptions (1.20) and (1.25), the rate of convergence in (1.21) of such adaptive predictors is usually inferior to that associated with recursive identification algorithms using matrix gains, as noted by Lai, Wei, and Zhang [18] who illustrated this point by the following simple example. Consider one-step ahead prediction in the ARX system

$$y_{n+1} = \alpha y_n + \beta u_n + \varepsilon_{n+1}, \tag{1.30}$$

where $|\alpha|<1$, the ε_n are independent normal random variables with mean 0 and variance $\sigma^2>0$, and the inputs u_n are also independent normal random variables with mean 0 and variance $\sigma_n^2 \sim n^{-2\gamma}$ for some $0<\gamma<\frac{1}{2}$ and such that $\{u_n\}$ and $\{\varepsilon_n\}$ are independent sequences. In this case, for the adaptive predictor $\hat{y}_{n+1}^G = \Theta_n'\Phi_n$ defined by the stochastic gradient algorithm (1.29) with $a=1$ and $\Phi_n=(y_n, u_n)'$, the convergence rate in (1.21) cannot be faster than $n^{-2\gamma}$ since

$$P\left\{\liminf_{n\to\infty} n^{-(1-2\gamma)} \sum_{i=1}^{n} (\tilde{y}_{i+1}-\hat{y}_{i+1}^G)^2>0\right\}>0. \tag{1.31}$$

On the other hand, the adaptive predictor $\hat{y}_{n+1}^{\mathrm{LS}}$ associated with the least squares estimate $(\sum_1^{n-1}\Phi_i\Phi_i')^{-1}\sum_1^{n-1}\Phi_i y_{i+1}$ of Θ satisfies

$$\limsup_{n\to\infty} \frac{\sum_{i=1}^n (\tilde{y}_{i+1}-\hat{y}_{i+1}^{\mathrm{LS}})^2}{\log n}<\infty \quad \text{a.s.}, \tag{1.32}$$

which implies that the convergence rate in (1.21) is $O(n^{-1}\log n)$ (cf. [18, p. 179]).

For the ARMAX system (1.1) with $C(q^{-1})\neq 1$, although the recursive estimator (1.13) has been called "approximate maximum likelihood" (AML), it does not arise from the maximization of the log-likelihood function, as in the off-line (nonrecursive) maximum likelihood estimator, when the ε_i are assumed to be normally distributed with mean 0 and variance σ^2. The recursive maximum likelihood estimator RML2, introduced by Åström and Söderström, replaces (1.13a) by

$$\Theta_n=\Theta_{n-1}+P_{n-1}\xi_{n-1}(y_n-\Theta_{n-1}'\Phi_{n-1}), \qquad P_n^{-1}=P_{n-1}^{-1}+\xi_n\xi_n', \tag{1.33a}$$

where letting $\Theta_n=(-\hat{a}_{n,1},\cdots,-\hat{a}_{n,p},\ \hat{b}_{n,1},\cdots,\hat{b}_{n,k},\ \hat{c}_{n,1},\cdots,\hat{c}_{n,h})'$, define ξ_n recursively by

$$\xi_n+\hat{c}_{n-1,1}\xi_{n-1}+\cdots+\hat{c}_{n-1,h}\xi_{n-h}=\Phi_n \tag{1.33b}$$

(cf. [2]). This algorithm is based on first replacing the derivative of the log-likelihood function by its linear approximation around the true parameter Θ and then replacing the unknown Θ by Θ_{n-1}, and would therefore lead to an asymptotically efficient estimator if Θ_{n-1} should converge to Θ. In § 5 we introduce an additional monitoring scheme to ensure that Θ_{n-1} is eventually close to Θ, and use the estended Lyapunov function (1.16) to analyze this modification of the RML2 algorithm, which we call the "monitored recursive maximum likelihood algorithm." We also extend in § 5 the RML2 algorithm (1.33) to the implicit system (1.9) and introduce an additional monitoring scheme to ensure that the recursive estimates θ_n generated by the algorithm are eventually close to the parameter vector θ of (1.9). By making use of extended stochastic Lyapunov functions, we are able to extend (1.32) to adaptive d-step ahead predictors based on θ_n.

In summary, the concept of extended stochastic Lyapunov functions provides a unified treatment of adpative predictors based on various recursive identification algorithms, using either the explicit or the implicit approach. In particular, we use this idea to improve in § 3 previous results on adaptive predictors based on extended least squares, to solve in § 4 an open problem in the literature concerning the global convergence of adaptive d-step ahead predictors based on the stochastic gradient algorithm without interlacing in the implicit approach, and to obtain in § 5 an analogue of (1.32) for the adaptive d-step ahead prediction problem by using the monitored recursive maximum likelihood algorithm that generalizes the least squares algorithm in (1.32). These results are of basic interest to adaptive control problems. In particular, while (1.21) leads to the self-optimizing property (1.22) of adaptive controllers based on the stochastic gradient algorithm, (1.33) and its extensions in § 5 lead to the much stronger property $\sum_{\Delta+1}^{n}(y_i-y_i^*-\eta_i)^2=O(\log n)$ almost surely for asymptotically efficient adaptive controllers.

2. Some preliminary lemmas. An important tool for the analysis of recursive identification and adaptive control algorithms is the following result from martingale theory.

LEMMA 1. *Let $\{\varepsilon_n\}$ be a martingale difference sequence with respect to an increasing sequence of σ-fields $\{\mathcal{F}_n\}$ such that* (1.3) *holds. Let z_n be an $\mathcal{F}_{n-1}$-measurable random variable for every n.*

(i) *$\sum_1^n z_i\varepsilon_i$ converges almost surely on $\{\sum_1^\infty z_i^2<\infty\}$, and for every $\eta>\frac{1}{2}$,*

$$\left(\sum_1^n z_i\varepsilon_i\right)\Big/\left(\sum_1^n z_i^2\right)^\eta\to 0\quad a.s.\ on\ \left\{\sum_1^\infty z_i^2=\infty\right\}.$$

Consequently,

$$\sum_1^n z_i\varepsilon_i=o\left(\sum_1^n z_i^2\right)+O(1)\quad a.s. \tag{2.1}$$

(ii) *$\sum_1^n|z_i|\varepsilon_i^2=O(\sum_1^n|z_i|)$ almost surely on $\{\sup_n|z_n|<\infty\}$. Moreover,*

$$\sum_1^n|z_i|\varepsilon_i^2=\sum_1^n|z_i|E(\varepsilon_i^2|\mathcal{F}_{i-1})+o\left(\sum_1^n|z_i|\right)\ on\ \left\{\sup_n|z_n|<\infty,\sum_1^\infty|z_n|=\infty\right\}. \tag{2.2}$$

(iii) *Let $T_1<T_2<\cdots$ be a sequence of stopping times (with respect to $\{\mathcal{F}_n\}$) such that $T_{j+1}-T_j\geqq d(\geqq 1)$. Let $\mathcal{G}_j=\mathcal{F}_{T_{j+1}}$ and let $w_j=|\sum_1^d\alpha_i\varepsilon_{T_j+i}|-E(|\sum_1^d\alpha_i\varepsilon_{T_j+i}||\mathcal{F}_{T_j})$, where $\alpha_1,\cdots,\alpha_d$ are constants. Then $\{w_j,\mathcal{G}_j,j\geqq 1\}$ is a martingale difference sequence with $\sup_j E(|w_j|^\alpha|\mathcal{G}_{j-1})<\infty$ almost surely.*

For the proof of parts (i) and (ii) of Lemma 1, see [19, p. 157], while part (iii) follows from that w_j is $\mathcal{G}_j$-measurable and that $E(w_j|\mathcal{F}_{T_j})=0$ and

$$E(|w_j|^\alpha|\mathcal{F}_{T_j})\leqq 2^\alpha\sum_{k=1}^\infty I_{\{T_j=k\}}E\left(\left|\sum_{i=1}^d\alpha_i\varepsilon_{k+i}\right|^\alpha\Big|\mathcal{F}_k\right).$$

While Lemma 1 is probabilistic in nature, Lemmas 2–5 below are algebraic. In particular, the algebraic identity (2.3) in Lemma 2 will be applied to analyze various recursive identification algorithms, in both the explicit and the implicit models. For a $\nu\times\nu$ matrix A, define $\|A\|=\sup_{\|x\|=1}\|Ax\|=\lambda_{\max}^{1/2}(A'A)$.

LEMMA 2. *Suppose that for $n\leqq t<m$, $\phi_t=(y_t,\cdots,y_{t-\nu+1},u_{t-\delta},\cdots u_{t-\delta-\kappa+1},\hat{y}_{t+d-1},\cdots,\hat{y}_{t+d-h})'$ and $C(q^{-1})(y_{t+d}-\eta_{t+d})=G(q^{-1})y_t+q^{-\delta}\Gamma(q^{-1})u_t$, where $\delta\geqq 0$, $d\geqq 1$, $\nu\geqq 1$, $\kappa\geqq 1$, and $C(q^{-1})=1+c_1q^{-1}+\cdots+c_hq^{-h}$, $G(q^{-1})=g_1+\cdots+g_\nu q^{-(\nu-1)}$ and $\Gamma(q^{-1})=\gamma_1+\cdots+\gamma_\kappa q^{-(\kappa-1)}$ are polynomials in the backward shift operator q^{-1}. Suppose that there exist $(\nu+\kappa+h)\times 1$ vectors θ_t such that $\hat{y}_{t+d}=\theta_t'\phi_t$ for $n\leqq t<m$. Then*

$$C(q^{-1})(y_s-\hat{y}_s-\eta_s)=-(\theta_{s-d}-\theta)'\phi_{s-d}\quad for\ m+d>s\geqq n+d, \tag{2.3}$$

where $\theta=(g_1,\cdots,g_\nu,\gamma_1,\cdots,\gamma_\kappa,-c_1,\cdots,-c_h)'$.

Proof. For $n\leqq t<m$,

$$\begin{aligned}C(q^{-1})&(y_{t+d}-\hat{y}_{t+d}-\eta_{t+d})\\&=C(q^{-1})(y_{t+d}-\eta_{t+d})-(C(q^{-1})-1)\hat{y}_{t+d}-\hat{y}_{t+d}\\&=(G(q^{-1})y_t+\Gamma(q^{-1})u_{t-\delta})-(c_1\hat{y}_{t+d-1}+\cdots+c_h\hat{y}_{t+d-h})-\theta_t'\phi_t\\&=\theta'\phi_t-\theta_t'\phi_t.\end{aligned}$$

□

LEMMA 3. *Let $\{D_n, n \geqq 0\}$ be a sequence of $L \times L$ real matrices such that $\sum_0^\infty \|D_n\| < \infty$ and let $D(z) = \sum_{n=0}^\infty D_n z^n$. Suppose that $D(e^{it}) + D'(e^{-it})$ is nonnegative definite for all $t \in [-\pi, \pi]$.*

(i) *Let $\{g_n, n \geqq 0\}$ be sequence of $L \times 1$ real vectors and let $f_n = \sum_{k=0}^n D_k g_{n-k}$. Then for any $N \geqq 0$, $\sum_{n=0}^N f_n' g_n \geqq 0$.*

(ii) *Suppose that $D_i = 0$ for all $i > h$. Let $M \geqq h$ and suppose that $f_n = \sum_{j=0}^h D_j g_{n-j}$ for $M \leqq n \leqq N$. Let $\{r_n, M \leqq n \leqq N\}$ be a nondecreasing sequence of positive numbers. Then*

$$\sum_{n=M}^{N} f_n' g_n / r_n \geqq \sum_{j=1}^{h} \sum_{t=0}^{h-j} g'_{M-1-t} D'_{j+t} g_{M-1+j} / r_{M-1+j}.$$

Proof. To prove (i), note that

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} \left(\sum_{n=0}^{N} g_n e^{-\mathrm{int}} \right)' \left(\frac{D(e^{it}) + D'(e^{-it})}{2} \right) \left(\sum_{n=0}^{N} g_n e^{int} \right) dt$$

$$= \frac{1}{2\pi} \sum_{k=0}^{\infty} \sum_{n=0}^{N} \sum_{m=0}^{N} g_n' D_k g_m \int_{-\pi}^{\pi} e^{i(m-n)t} e^{ikt} \, dt$$

$$= \sum_{k=0}^{N} \sum_{\substack{0 \leqq m \leqq n \leqq N \\ m-n=-k}} g_n' D_k g_m = \sum_{n=0}^{N} g_n' \sum_{k=0}^{n} D_k g_{n-k} = \sum_{n=0}^{N} g_n' f_n.$$

Since $D(e^{it}) + D'(e^{-it})$ is nonnegative definite, (i) follows.

To prove (ii), let $g_n^* = g_n$ if $n \geqq M$ and let $g_n^* = 0$ otherwise. Let $f_n^* = \sum_{j=0}^h D_j g_{n-j}^*$. Then $f_n^* = f_n$ for $n \geqq M + h$ and $f_n - f_n^* = \sum_{m=n-h}^{M-1} D_{n-m} g_m$ for $M \leqq n < M + h$. Therefore

$$\sum_{n=M}^{N} g_n' f_n / r_n = \sum_{n=M}^{N} g_n^{*\prime} f_n^* / r_n + \sum_{j=1}^{h} g'_{M-1+j} \left(\sum_{k=j}^{h} D_k g_{M-1+j-k} \right) \Big/ r_{M-1+j}.$$

Let $S_n = \sum_{m=0}^n g_m^{*\prime} f_m^*$. By (i), $S_n \geqq 0$ for all $n \leqq N$. Since $g_n^* = 0$ for $n < M$, $S_{M-1} = 0$. Summation by parts gives

$$\sum_{n=M}^{N} g_n^{*\prime} f_n^* / r_n = r_N^{-1} S_N + \sum_{n=M}^{N-1} (r_n^{-1} - r_{n+1}^{-1}) S_n \geqq 0,$$

noting that $r_n^{-1} \geqq r_{n+1}^{-1} > 0$. Hence the desired conclusion follows. □

LEMMA 4. *Let $x_1, x_2, \cdots$ be $\nu \times 1$ vectors and let $A_n = A_{n-1} + x_n x_n' + \rho_n I$, where ρ_n are nonnegative scalars, I is the $\nu \times \nu$ identity matrix, and A_0 is a symmetric, positive-definite $\nu \times \nu$ matrix.*

(i) *If $\lim_{n\to\infty} \lambda_{\max}(A_n) < \infty$, then $\sum_{i=1}^\infty x_i' A_i^{-1} x_i < \infty$. If $\lim_{n\to\infty} \lambda_{\max}(A_n) = \infty$, then*

$$\sum_{i=1}^{n} x_i' A_i^{-1} x_i \leqq (1 + o(1)) \log \det A_n. \tag{2.4}$$

(ii) *Suppose that $\sup_n \rho_n < \infty$ and that $\lambda_{\min}(A_n) \to \infty$ and $x_n' A_n^{-1} x_n \to 0$ as $n \to \infty$. Then for every fixed $r = 0, \pm 1, \pm 2, \cdots$,*

$$x_n' A_{n+r}^{-1} x_n \sim x_n' A_n^{-1} x_n \quad \textit{as } n \to \infty. \tag{2.5}$$

Proof. By Lemma 2(i) of [19],

$$x_i' A_i^{-1} x_i = (|A_i| - |A_{i-1} + \rho_i I|) / |A_i|. \tag{2.6}$$

Since $|A_{i-1} + \rho_i I| \geqq |A_{i-1}|$, it follows from (2.6) that $x_i' A_i^{-1} x_i \leqq (|A_i| - |A_{i-1}|)/|A_i|$ and therefore (i) follows by the same argument as that used to prove Lemma 2(ii) of [19].

To prove (ii), we first show that

$$x_n' A_{n-1}^{-1} x_n \sim x_n' A_n^{-1} x_n \quad \text{as } n \to \infty. \tag{2.7}$$

By the matrix inversion lemma (cf. [3, p. 824]),

$$A_n^{-1} = (A_{n-1}+\rho_n I)^{-1} - \frac{(A_{n-1}+\rho_n I)^{-1} x_n x_n' (A_{n-1}+\rho_n I)^{-1}}{1 + x_n'(A_{n-1}+\rho_n I)^{-1} x_n},$$

and therefore

$$x_n' A_n^{-1} x_n = x_n'(A_{n-1}+\rho_n I)^{-1} x_n / \{1 + x_n'(A_{n-1}+\rho_n I)^{-1} x_n\}. \tag{2.8}$$

Since $(A_{n-1}+\rho_n I)^{-1} = A_{n-1}^{-1/2}(I+\rho_n A_{n-1}^{-1})^{-1} A_{n-1}^{-1/2} = A_{n-1}^{-1} + A_{n-1}^{-1/2} B_n A_{n-1}^{-1/2}$, where $B_n = \sum_{i=1}^{\infty} (-\rho_n A_{n-1}^{-1})^i$, and since $\rho_n \lambda_{\max}(A_{n-1}^{-1}) \to 0$ (implying that $\lambda_{\max}(B_n) \to 0$), it then follows that

$$x_n'(A_{n-1}+\rho_n I)^{-1} x_n = x_n' A_{n-1}^{-1} x_n + o(\|A_{n-1}^{-1/2} x_n\|^2) = (1+o(1)) x_n' A_{n-1}^{-1} x_n. \tag{2.9}$$

From (2.8) and (2.9), (2.7) follows.

We next show by induction that for $r = 1, 2, \cdots$,

$$x_n' A_{n-r}^{-1} x_n \sim x_n' A_n^{-1} x_n \quad \text{as } n \to \infty. \tag{2.10}$$

Note that (2.10) reduces to (2.7) when $r = 1$. Suppose that (2.10) holds for $1 \leqq r \leqq s-1$. Since $A_{n-1} = A_{n-s} + \sum_{j=1}^{s-1} (x_{n-j} x_{n-j}' + \rho_{n-j} I)$,

$$\begin{aligned} A_{n-1}^{-1} &= A_{n-s}^{-1/2} \left\{ I + \sum_{j=1}^{s-1} (A_{n-s}^{-1/2} x_{n-j} x_{n-j}' A_{n-s}^{-1/2} + \rho_{n-j} A_{n-s}^{-1}) \right\}^{-1} A_{n-s}^{-1/2} \\ &= A_{n-s}^{-1} + A_{n-s}^{-1/2} C_n A_{n-s}^{-1/2}, \end{aligned} \tag{2.11}$$

where

$$C_n = \sum_{i=1}^{\infty} \left\{ -\sum_{j=1}^{s-1} (A_{n-s}^{-1/2} x_{n-s+j} x_{n-s+j}' A_{n-s}^{-1/2} + \rho_{n-s+j} A_{n-s}^{-1}) \right\}^i. \tag{2.12}$$

Noting that $\|C\| = \lambda_{\max}(C) \leqq \operatorname{tr}(C)$ if C is symmetric and nonnegative definite, we have

$$\sum_{j=1}^{s-1} \|A_{n-s}^{-1/2} x_{n-s+j} x_{n-s+j}' A_{n-s}^{-1/2}\| \leqq \sum_{j=1}^{s-1} x_{n-s+j}' A_{n-s}^{-1} x_{n-s+j} \to 0, \tag{2.13}$$

since (2.10) holds for $1 \leqq r \leqq s-1$. From (2.13) and the fact that $\rho_{n-s+j} \lambda_{\max}(A_{n-s}^{-1}) = \rho_{n-s+j} / \lambda_{\min}(A_{n-s}) \to 0$, it follows that $\|C_n\| \to 0$, and therefore by (2.11),

$$x_n' A_{n-1}^{-1} x_n = x_n' A_{n-s}^{-1} x_n + o(x_n' A_{n-s}^{-1} x_n). \tag{2.14}$$

In view of (2.14) and (2.7), (2.10) holds for $r = s$.

For $r \geqq 1$, since $A_{n+r} = A_n + \sum_{j=1}^{r} (x_{n+j} x_{n+j}' + \rho_{n+j} I)$, we can make use of (2.10) and the same argument as in (2.11)-(2.14) to show that $x_n' A_{n+r}^{-1} x_n = x_n' A_n^{-1} x_n + o(x_n' A_n^{-1} x_n)$. □

LEMMA 5. *Suppose that the polynomial* $C(z) = 1 + c_1 z + \cdots + c_h z^h$ *is stable. For* $j = 1, \cdots, h$, *let* $\{c_{n,j}\}$ *be a sequence of numbers such that* $\lim_{n\to\infty} c_{n,j} = c_j$, *and let* $C_n(q^{-1}) = 1 + c_{n,1} q^{-1} + \cdots + c_{n,h} q^{-h}$.

(i) *Suppose that* ξ_n *and* ϕ_n *are* $L \times 1$ *vectors such that* $C_n(q^{-1})\xi_n = \phi_n$. *Then there exist* $K > 0$ *and* $0 < \rho < 1$ *such that for all* $t > m$

$$\|\xi_t\| \leqq K \left\{ \sum_{i=0}^{t-m-1} \rho^i \|\phi_{t-i}\| + \rho^{t-m} \sum_{r=0}^{h-1} \|\xi_{m-r}\| \right\}. \tag{2.15}$$

Consequently, there exists $K' > 0$ *such that for all* $n > m$,

$$\sum_{t=m+1}^{n} \|\xi_t\|^2 \leqq K' \left\{ \max_{0 \leqq r \leqq h-1} \|\xi_{m-r}\|^2 + \sum_{t=m+1}^{n} \|\phi_t\|^2 \right\}. \tag{2.16}$$

(ii) *Let* $G(q^{-1}) = g_1 + \cdots + g_p q^{-p+1}$, $\Gamma(q^{-1}) = \gamma_1 + \cdots + \gamma_k q^{-k+1}$, $G_n(q^{-1}) = g_{n,1} + \cdots + g_{n,p} q^{-p+1}$, $\Gamma_n(q^{-1}) = \gamma_{n,1} + \cdots + \gamma_{n,k} q^{-k+1}$, *where* $\lim_{n\to\infty} g_{n,j} = g_j$ *and* $\lim_{n\to\infty} \gamma_{n,j} = \gamma_j$ *for every j. Suppose that*

$$C(q^{-1})\tilde{y}_{n+d} = G(q^{-1})y_n + \Gamma(q^{-1})u_n, \qquad C_n(q^{-1})\hat{y}_{n+d} = G_n(q^{-1})y_n + \Gamma_n(q^{-1})u_n. \tag{2.17}$$

Then $\sum_{i=1}^n (\tilde{y}_{i+d} - \hat{y}_{i+d})^2 = o(\sum_{i=1}^n (y_i^2 + u_i^2))$.

Proof. For (i), see [19, pp. 161-162] and [9, p. 904]. To prove (ii), note that by (2.17),

$$\begin{aligned} C_n(q^{-1})(\hat{y}_{n+d} - \tilde{y}_{n+d}) &= C_n(q^{-1})\hat{y}_{n+d} - C(q^{-1})\tilde{y}_{n+d} - (C_n(q^{-1}) - C(q^{-1}))\tilde{y}_{n+d} \\ &= (G_n(q^{-1}) - G(q^{-1}))y_n + (\Gamma_n(q^{-1}) - \Gamma(q^{-1}))u_n \\ &\quad - (C_n(q^{-1}) - C(q^{-1}))\tilde{y}_{n+d}. \end{aligned} \tag{2.18}$$

Since $G_n - G \to 0$, $\Gamma_n - \Gamma \to 0$, and $C_n - C \to 0$, it follows from (2.18) and part (i) of the lemma that

$$\sum_{i=1}^n (\tilde{y}_{i+d} - \hat{y}_{i+d})^2 = o\left(\sum_{i=1}^n (y_i^2 + u_i^2 + \tilde{y}_{i+d}^2)\right). \tag{2.19}$$

Again by (2.17) and (i),

$$\sum_{i=1}^n \tilde{y}_{i+d}^2 = O\left(\sum_{i=1}^n (y_i^2 + u_i^2)\right). \tag{2.20}$$

From (2.19) and (2.20), the desired conclusion follows. □

3. Extended least squares and the associated adaptive predictors. To begin with, consider the model (1.1) with $C(q^{-1}) = 1$, i.e., the ARX model. In this case, the AML algorithm (1.13) reduces to the usual least squares estimator, for which (1.13b) becomes $\Phi_n = (y_n, \cdots, y_{n-p+1}, u_{n-\Delta+1}, \cdots, u_{n-\Delta-k+2})' = \Psi_n$. As shown in [18] and [19], the least squares one-step ahead predictors $\hat{y}_{n+1} = \Theta_n' \Phi_n$ satisfy

$$\sum_{n=1}^N (\tilde{y}_{n+1} - \hat{y}_{n+1})^2 I_{\{\Phi_n' P_n \Phi_n \leq \delta\}} = O(\log \det P_N^{-1}) \quad \text{for every } 0 < \delta < 1, \tag{3.1}$$

where $\tilde{y}_{n+1} = \Theta' \Psi_n$ is the optimal one-step ahead predictor of y_{n+1} assuming knowledge of the parameter vector $\Theta = (-a_1, \cdots, -a_p, b_1, \cdots, b_k)'$ and $P_n^{-1} = P_{n-1}^{-1} + \Phi_n \Phi_n'$ as in (1.13a). When the sample mean square boundedness assumption (1.20) holds for the input-output data, (3.1) implies that

$$\sum_{i=1}^N (\tilde{y}_{n+1} - \hat{y}_{n+1})^2 = O(\log N) \quad \text{a.s. on } \{\limsup_{n\to\infty} \Phi_n' P_n \Phi_n < 1\}. \tag{3.2}$$

For general $C(z)$ satisfying the positive real condition (1.15), Lai and Wei [9] showed that (3.1) still holds for the adaptive one-step ahead predictors $\hat{y}_{n+1} = \Theta_n' \Phi_n$ based on the AML algorithm Θ_n and the pseudoregression vector Φ_n defined in (1.13). It is also shown in [9] that

$$\sum_{n=1}^N \|\Phi_n - \Psi_n\|^2 = O\left(\log\left(2 + \sum_{n=1}^N \|\Phi_n\|^2\right)\right) \quad \text{a.s.} \tag{3.3}$$

If (1.20) holds, then since $\sum_{n=1}^N \varepsilon_n^2 = O(N)$ almost surely by Lemma 1(ii), $\sum_{n=1}^N \|\Psi_n\|^2 = O(N)$ almost surely and therefore $\sum_{n=1}^N \|\Phi_n\|^2 = O(N)$ almost surely by (3.3). Hence, under assumptions (1.15) and (1.20), (3.2) still holds for the adaptive predictor $\hat{y}_{n+1}$ based on the AML algorithm.

In [9], (3.1) was obtained from an analysis of the recursive inequalities for the extended stochastic Lyapunov function (1.16) that was also used to study the consistency of Θ_n. Earlier, Solo [8] used these recursive inequalities to transform (1.16) into a nonnegative almost supermartingale and thereby applies the martingale convergence theorem to establish the strong consistency of Θ_n under assumptions (1.14) and (1.15). There is, however, a gap in Solo's proof, as noted by Zhang [13]. Specifically, Solo's proof made use of claim (A6) in Appendix I of [8] that for $\nu \times 1$ vectors x_i, if $\sum_1^n \|x_i\|^2 = O(n)$ and $\sum_1^n x_i x_i'$ is nonsingular, then $x_n'(\sum_1^n x_i x_i')^{-1} x_n \to 0$. Zhang [13] found an error in the proof of (A6) and concluded that (A6) is "questionable." In fact, (A6) turns out to be false, as can be seen from the following example in the scalar case $\nu = 1$. Let $J = \{2, 2^2, 2^3, \cdots\}$ and let $x_n = 1$ if $n \notin J$ and $x_n = n^{1/2}$ if $n \in J$. Then $n \leqq \sum_1^n x_i^2 \leqq n + \sum_{i:2^i \leqq n} 2^i \leqq 3n$ for all n, and $x_n^2/\sum_1^n x_i^2 \geqq \frac{1}{3}$ for $n \in J$, violating (A6).

For d-step ahead prediction, Sin, Goodwin, and Bitmead [12] proposed an extension of the AML algorithm to construct adaptive predictors using the following implicit approach. Instead of working with (1.5), they introduced a further reparametrization to facilitate the analysis of the AML algorithm that directly estimates the parameters of this reparametrized model. Applying the division algorithm, they wrote

$$1 = \bar{F}(z)C(z) + z^d \bar{G}(z), \tag{3.4}$$

where $\bar{F}(z) = 1 + \bar{f}_1 z + \cdots + \bar{f}_{d-1} z^{d-1}$ and $\bar{G}(z) = \bar{g}_1 + \bar{g}_2 z + \cdots + \bar{g}_h z^{h-1}$. Let

$$\bar{C}(z) = 1 - z^d \bar{G}(z) = 1 - \bar{g}_1 z^d - \cdots - \bar{g}_h z^{d+h-1}. \tag{3.5}$$

From (3.4), $\bar{C}(z) = \bar{F}(z)C(z)$. Multiplying (1.5) by $\bar{F}(q^{-1})$ gives

$$\begin{aligned} &\bar{C}(q^{-1})\{y_{n+d} - F(q^{-1})\varepsilon_{n+d}\} \\ &\quad = \bar{F}(q^{-1})G(q^{-1})y_n + q^{-(\Delta-d)}\bar{F}(q^{-1})F(q^{-1})B(q^{-1})u_n. \end{aligned} \tag{3.6}$$

Therefore, analogous to (1.9), system (1.1) can be written in the prediction form

$$\begin{aligned} &y_{n+d} = \bar{\theta}'\bar{\psi}_n + \eta_{n+d} \quad \text{where } \eta_{n+d} = F(q^{-1})\varepsilon_{n+d} = y_{n+d} - \tilde{y}_{n+d}, \\ &\bar{\psi}_n = (y_n, \cdots, y_{n-p(d)-d+2}, u_{n-\Delta+d}, \cdots, u_{n-k-\Delta-d+3}, \tilde{y}_n, \cdots, \tilde{y}_{n-h+1})', \\ &\bar{\theta} = (g_1, g_1\bar{f}_1 + g_2, \cdots, \bar{f}_{d-1} f_{d-1} b_k, \bar{g}_1, \cdots, \bar{g}_h)'. \end{aligned} \tag{3.7}$$

In analogy with (1.13), Sin, Goodwin, and Bitmead [12] introduced the following extended least squares algorithm to estimate $\bar{\theta}$:

$$\bar{\theta}_n = \bar{\theta}_{n-d} + P_{n-d}\bar{\phi}_{n-d}(y_n - \bar{\theta}'_{n-d}\bar{\phi}_{n-d}), \tag{3.8a}$$

$$P_n^{-1} = P_{n-d}^{-1} + \bar{\phi}_n \bar{\phi}'_n, \tag{3.8b}$$

$$\begin{aligned} \bar{\phi}_n = (&y_n, \cdots, y_{n-p(d)-d+2}, u_{n-\Delta+d}, \cdots, u_{n-k-\Delta-d+3}, \\ &\bar{\theta}'_n\bar{\phi}_{n-d}, \cdots, \bar{\theta}'_{n-h+1}\bar{\phi}_{n-h+1-d})'. \end{aligned} \tag{3.8c}$$

Thus, (3.8) is an AML-type algorithm which replaces the $\tilde{y}_i$ in $\bar{\psi}_n$ by the "a posteriori" predictor $\bar{\theta}'_i\bar{\phi}_{i-d}$. Note also that (3.8) can be regarded as "interlacing" d unit-delay-type recursions for $\bar{\theta}_{j+dt}$ $(j = 0, \cdots, d-1)$.

The extended least squares d-step ahead predictor is

$$\hat{y}_{n+d} = \bar{\theta}'_n \bar{\phi}_n, \tag{3.9}$$

where $\bar{\theta}_n$ and $\bar{\phi}_n$ are given by (3.8). Assuming (1.20) and

$$\min_{|z|=1} \operatorname{Re}\left(\frac{1}{\bar{C}(z)} - \frac{1}{2}\right) > 0, \tag{3.10}$$

Sin, Goodwin, and Bitmead [12] modified Solo's [8] argument to prove that $n^{-1}\sum_1^n (\tilde{y}_{i+d}-\hat{y}_{i+d})^2 \to 0$ almost surely, i.e., the adaptive predictors (3.9) are globally convergent. As noted by Zhang [13], their proof uses Solo's [8] result (A6) to conclude that $\bar{\phi}'_n P_n \bar{\phi}_n \to 0$ almost surely. Since (A6) has been shown to be invalid, their proof only gives that under (1.20) and (3.10),

$$n^{-1}\sum_{i=1}^{n} (\tilde{y}_{i+d}-\hat{y}_{i+d})^2 \to 0 \quad \text{a.s.} \qquad \text{if } \lim_{n\to\infty} \bar{\phi}'_n P_n \bar{\phi}_n = 0 \quad \text{a.s.} \tag{3.11}$$

Modifying the proof of Theorem 1 of Lai and Wei [9] (instead of Solo's [8] arguments) and combining it with algebraic details similar to those provided by [12] leads to the following analogue of (3.2), which is considerably stronger than (3.11).

THEOREM 1. *Suppose that the random disturbances* ε_n *in the linear stochastic system* (1.1) *satisfy* (1.3) *and that the positive real assumption* (3.10) *holds. Consider the extended least squares algorithm* $\bar{\theta}_n$, *defined by* (3.8) *for the implicit model* (3.7), *and its associated adaptive d-step ahead predictor* (3.9). *Then for every* $0<\delta<1$,

$$\sum_{i=1}^{n} (\tilde{y}_{i+d}-\hat{y}_{i+d})^2 I_{\{\bar{\phi}'_i P_i \bar{\phi}_i \leqq \delta\}} = O(\log \det P_n^{-1}) \quad \textit{a.s.}, \tag{3.12}$$

$$\sum_{i=1}^{n} \|\bar{\phi}_i - \bar{\psi}_i\|^2 = O\left(\log\left(2+\sum_{i=1}^{n} \|\bar{\phi}_i\|^2\right)\right) \quad \textit{a.s.}, \tag{3.13}$$

where $\bar{\psi}_i$ *is defined in* (3.7). *Furthermore, if* (1.20) *also holds, then* $\sum_1^n \|\bar{\phi}_i\|^2 = O(n)$ *almost surely and*

$$\sum_{i=1}^{n} (\tilde{y}_{i+d}-\hat{y}_{i+d})^2 = O(\log n) \quad \textit{a.s. on } \{\limsup_{n\to\infty} \bar{\phi}'_n P_n \bar{\phi}_n < 1\}. \tag{3.14}$$

Proof. Let $f_n = (\bar{\theta}-\bar{\theta}_n)'\bar{\phi}_{n-d}$, $e_n = y_n - \bar{\theta}'_{n-d}\bar{\phi}_{n-d}$, $\hat{\eta}_n = y_n - \bar{\theta}'_n \bar{\phi}_{n-d}$, $w_n = \hat{\eta}_n - \eta_n$. By (3.16) of [12], $\bar{C}(q^{-1})w_n = f_n$. Define

$$Q_n = (\bar{\theta}_n - \bar{\theta})' P_{n-d}^{-1} (\bar{\theta}_n - \bar{\theta}). \tag{3.15}$$

Fix $j \in \{0, \cdots, d-1\}$ and let $a_{t,j} = a_{j+dt}$ for a sequence of vectors (matrices, scalars) a_n, so that $\bar{\theta}_{t,j} = \bar{\theta}_{j+dt}$, $P_{t,j} = P_{j+dt}$, etc. By (3.13) of [12],

$$\hat{\eta}_{t,j} = (1 - \bar{\phi}'_{t-1,j} P_{t-1,j} \bar{\phi}_{t-1,j}) e_{t,j}. \tag{3.16}$$

As shown in [12, p. 1163], it follows from (3.8), (3.15), and (3.16) that

$$Q_{t,j} = Q_{t-1,j} + f_{t,j}^2 - 2f_{t,j}\hat{\eta}_{t,j} - \bar{\phi}'_{t-1,j}P_{t-1,j}\bar{\phi}_{t-1,j}(1-\bar{\phi}'_{t-1,j}P_{t-1,j}\bar{\phi}_{t-1,j})e_{t,j}^2. \tag{3.17}$$

By (3.8a),

$$-f_{t,j} = (\bar{\theta}_{t-1,j}-\bar{\theta})'\bar{\phi}_{t-1,j} + \bar{\phi}'_{t-1,j}P_{t-1,j}\bar{\phi}_{t-1,j}(e_{t,j}-\eta_{t,j}) + \bar{\phi}'_{t-1,j}P_{t-1,j}\bar{\phi}_{t-1,j}\eta_{t,j}. \tag{3.18}$$

Since $\hat{\eta}_{t,j} = w_{t,j} + \eta_{t,j}$, it follows from (3.17) and (3.18) that

$$\begin{aligned} Q_{t,j} = {} & Q_{t-1,j} - f_{t,j}(2w_{t,j} - f_{t,j}) \\ & + 2\eta_{t,j}\{(\bar{\theta}_{t-1,j}-\bar{\theta})'\bar{\phi}_{t-1,j} + \bar{\phi}'_{t-1,j}P_{t-1,j}\bar{\phi}_{t-1,j}(e_{t,j}-\eta_{t,j})\} \\ & + 2\bar{\phi}'_{t-1,j}P_{t-1,j}\bar{\phi}_{t-1,j}\eta_{t,j}^2 - \bar{\phi}'_{t-1,j}P_{t-1,j}\bar{\phi}_{t-1,j}(1-\bar{\phi}'_{t-1,j}P_{t-1,j}\bar{\phi}_{t-1,j})e_{t,j}^2, \end{aligned} \tag{3.19}$$

which is of the same form as (24) in [9]. Let $g_i = w_i - f_i/2$. For $n = dT + \nu$ with $0 \leqq \nu < d$, we obtain by summing (3.19) that

$$Q_n + Q_{n-1} + \cdots + Q_{n-d+1} = Q_0 + \cdots + Q_{d-1} - 2 \sum_{i=d}^{n} f_i g_i$$

$$- \left(\sum_{j=0}^{\nu} \sum_{t=1}^{T} + \sum_{j=\nu+1}^{d-1} \sum_{t=1}^{T-1} \right) [2\eta_{t,j}\{(\bar{\theta}_{t-1,j} - \bar{\theta})' \bar{\phi}_{t-1,j} + \bar{\phi}'_{t-1,j} P_{t-1,j} \bar{\phi}_{t-1,j}(e_{t,j} - \eta_{t,j})\} \tag{3.20}$$

$$+ 2\bar{\phi}'_{t-1,j} P_{t-1,j} \bar{\phi}_{t-1,j} \eta^2_{t,j} - \bar{\phi}'_{t-1,j} P_{t-1,j} \bar{\phi}_{t-1,j} (1 - \bar{\phi}'_{t-1,j} P_{t-1,j} \bar{\phi}_{t-1,j}) e^2_{t,j}].$$

Since $w_i = (1/\bar{C}(q^{-1}))f_i$ by the stability of $\bar{C}(z)$, $g_i = (1/\bar{C}(q^{-1}) - \frac{1}{2})f_i$ and therefore by the positive real assumption (3.10) and Lemmas 3 and 5(i), there exist $\delta > 0$ (for which $1/\bar{C}(z) - \frac{1}{2} - \delta$ is positive real) and $K > 0$ such that

$$\sum_{i=d}^{n} f_i g_i + K \geqq \delta \sum_{i=d}^{n} f_i^2, \quad \sum_{i=d}^{n} f_i g_i + K \geqq \delta \sum_{i=d}^{n} f_i^2 \quad \text{for all } n \geqq d, \tag{3.21}$$

analogous to (25) in [9]. Note also that (3.16) and (3.18) are analogous to (26) of [9]. The rest of the proof of (3.12) and (3.13) is therefore the same as that of Theorem 1 of [9], where the only references to Solo [8] are related to (24) and (25) of [9], and (A6) of [8] is never used.

Suppose that (1.20) holds. Since $\sum_1^n \varepsilon_i^2 = O(n)$ almost surely by Lemma 1(ii), $\sum_1^n \eta_i^2 = O(n)$ almost surely. Hence $\sum_1^n \tilde{y}_i^2 = \sum_1^n (y_i - \eta_i)^2 \leqq 2(\sum_1^n y_i^2 + \sum_1^n \eta_i^2) = O(n)$ almost surely. Therefore $\sum_1^n \|\bar{\psi}_i\|^2 = O(n)$ almost surely, which together with (3.13) gives that

$$\sum_{i=1}^{n} \|\bar{\phi}_i\|^2 \leqq 2 \sum_{i=1}^{n} \|\bar{\psi}_i\|^2 + 2 \sum_{i=1}^{n} \|\bar{\phi}_i - \bar{\psi}_i\|^2 = O\left(n + \log \left(2 + \sum_{i=1}^{n} \|\bar{\phi}_i\|^2 \right) \right) \quad \text{a.s.,}$$

implying that $\sum_1^n \|\bar{\phi}_i\|^2 = O(n)$ almost surely. Hence

$$\log \det P_n^{-1} = O\left(\log \left(2 + \sum_1^n \|\bar{\phi}_i\|^2 \right) \right) = O(\log n)$$

almost surely, and therefore (3.14) follows from (3.12). □

Remark. Unlike the preceding arguments, the proof of Sin, Goodwin, and Bitmead [12] uses (3.17) to show that

$$\begin{aligned} &E[n^{-1}Q_n | \mathcal{F}_{n-d}] + E[n^{-1}S_n | \mathcal{F}_{n-d}] \\ &\quad \leqq (n-d)^{-1} Q_{n-d} + E[(n-1)^{-1} S_{n-1} | \mathcal{F}_{n-d}] \\ &\qquad - \rho n^{-1} E[f_n^2 | \mathcal{F}_{n-d}] + 2n^{-1} \bar{\phi}'_{n-d} P_{n-d} \bar{\phi}_{n-d} E[\eta_n^2 | \mathcal{F}_{n-d}], \end{aligned} \tag{3.22}$$

where $S_n = 2\sum_{i=d}^{n} f_i(g_i - \frac{1}{2}\rho f_i) + K$, and ρ and K are chosen by (3.10) so that $1/\bar{C}(z) - (1+\rho)/2$ is positive real and $S_n \geqq 0$ for all $n \geqq d$. Summing (3.22) from d to N and taking expectations, a crucial step in their argument is to prove that

$$E\left(\sum_{n=d}^{\infty} n^{-1} \bar{\phi}'_{n-d} P_{n-d} \bar{\phi}_{n-d} \right) < \infty, \tag{3.23}$$

which is then used to show that $E(\sum_{n=d}^{\infty} n^{-1} f_n^2) < \infty$, from which it follows that $\sum_d^{\infty} n^{-1} f_n^2 < \infty$ almost surely. Sin, Goodwin, and Bitmead [12] prove (3.23) by first showing that $\sum_d^{\infty} n^{-1} \bar{\phi}'_{n-d} P_{n-d} \bar{\phi}_{n-d} < \infty$ almost surely and then applying the monotone convergence theorem to conclude that (3.23) holds. However, this application of the monotone convergence theorem is invalid, and we cannot conclude from the almost sure finiteness of $Z \triangleq \sum_1^{\infty} n^{-1}(z_n / \sum_1^n z_i)$ that $EZ < \infty$ for nonnegative random variables z_n. In particular, letting $z_n = u_{n-\Delta}^2 = \bar{\phi}_{n-d}^2$ in the special case $d = \Delta = 1$ and $A(q^{-1}) = B(q^{-1}) = C(q^{-1}) = 1$, consider the following counterexample. Let $v_0, v_1, \cdots$ be independent exponential random variables with mean 1. Define $z_0 = v_0$, $z_n = nI_{\{\sum_1^{n-1} z_i < 2n\}} + v_n I_{\{\sum_1^{n-1} z_i \geqq 2n\}}$ for $n \geqq 1$. Then a standard argument shows that $\liminf_{n\to\infty} E(z_n/\sum_1^n z_i) > 0$ and that $P\{n/2 \leqq \sum_1^n z_i \leqq 3n$ for all large $n\} = 1$. Therefore, although $Z = \sum_1^\infty n^{-1}(z_n/\sum_1^n z_i) \leqq 3\sum_1^\infty z_n/(\sum_1^n z_i)^2 + O(1) < \infty$ almost surely, $EZ = \sum_1^\infty n^{-1}E(z_n/\sum_1^n z_i) = \infty$.

Theorem 1 suggests that if we use the extended least squares predictor $\hat{y}_{n+d} = \bar{\theta}'_n \bar{\phi}_n$ to predict y_{n+d} whenever $\bar{\phi}'_n P_n \bar{\phi}_n \leqq \delta$ (with $0 < \delta < 1$) and use some other predictor when this condition is violated, then we may be able to achieve a good overall performance. This idea will be further developed in Corollaries 2 and 3 of § 4 where we will use the stochastic gradient algorithm to give an adaptive predictor of y_{n+d} when $\bar{\phi}_n P'_n \bar{\phi}_n > \delta$.

4. The stochastic gradient algorithm and some extensions. To begin with, consider the stochastic gradient algorithm (1.29) that estimates the parameter vector (1.11) of the explicit system (1.1). As an application of Lemma 1 to the extended stochastic Lyapunov function $Q_n = \|\Theta_n - \Theta\|^2$, we prove the following corollary, part (ii) of which deals with a modification of (1.29) that constrains by projection the estimator to lie inside some convex region. Such constrained algorithms have been studied by Ljung [5], [6] and Kushner and Clark [20] by the ODE method.

COROLLARY 1. *Suppose that* $\min_{|z|=1} \operatorname{Re}\{C(z) - a/2\} > 0$ *and that the random disturbances* ε_n *in the linear stochastic system* (1.1) *satisfy* (1.3).

(i) *For the stochastic gradient algorithm* (1.29),

$$\limsup_{n\to\infty} \|\Theta_n\| < \infty \quad a.s., \tag{4.1}$$

$$\sum_{n=1}^{\infty} (e_n - \varepsilon_n)^2 / r_n < \infty \quad a.s., \tag{4.2}$$

$$\sum_{n=1}^{\infty} \|\Theta_n - \Theta_{n-1}\|^2 < \infty \quad a.s., \tag{4.3}$$

where e_n *and* r_n *are as defined in* (1.29b) *and* (1.29c).

(ii) *Let* $\{D_n\}$ *be a sequence of closed convex regions in* R^{p+k+h} *such that* $\Theta \in D_n$ *for all large n. Let* $\Pi_D(x)$ *denote the Euclidean projection of* x *into* D, *i.e.*, $\|x - \Pi_D(x)\| = \min\{\|x - y\|: y \in D\}$. *Suppose that we modify* (1.29a) *defining the stochastic gradient algorithm as*

$$\Theta_n = \Pi_{D_n}(\Theta_{n-1} + ar_{n-1}^{-1} e_n \Phi_{n-1}). \tag{4.4}$$

Then (4.1) *and* (4.2) *still hold. If, furthermore,* $D_n \supset D_{n-1}$ *for all large n, then* (4.3) *still holds.*

Proof. (i) Let $Q_i = \|\Theta_i - \Theta\|^2$. From (1.29a) it follows that

$$Q_i = Q_{i-1} + 2ar_{i-1}^{-1} e_i \Phi'_{i-1}(\Theta_{i-1} - \Theta) + a^2 r_{i-1}^{-2} e_i^2 \|\Phi_{i-1}\|^2. \tag{4.5}$$

Replacing e_i in (4.5) by $(e_i - \varepsilon_i) + \varepsilon_i$ and summing (4.5) over i give that for $n > h$,

$$\begin{aligned} Q_n = Q_h &+ 2a \sum_{i=h+1}^{n} \{r_{i-1}^{-1}(e_i - \varepsilon_i)\Phi'_{i-1}(\Theta_{i-1} - \Theta) + (a/2) r_{i-1}^{-2}(e_i - \varepsilon_i)^2 \|\Phi_{i-1}\|^2\} \\ &+ a^2 \sum_{i=h+1}^{n} r_{i-1}^{-2} \|\Phi_{i-1}\|^2 \varepsilon_i^2 + 2a \sum_{i=h+1}^{n} \varepsilon_i \\ &\times \{r_{i-1}^{-1}\Phi'_{i-1}(\Theta_{i-1} - \Theta) + ar_{i-1}^{-2}(e_i - \varepsilon_i)\|\Phi_{i-1}\|^2\}. \end{aligned} \tag{4.6}$$

Since $e_i-\varepsilon_i=(y_i-\varepsilon_i)-\Theta'_{i-1}\Phi_{i-1}$ is $\mathscr{F}_{i-1}$-measurable, it follows from (2.1) that

$$\begin{aligned}&\sum_{i=h+1}^{n}\varepsilon_i\{r_{i-1}^{-1}\Phi'_{i-1}(\Theta_{i-1}-\Theta)+ar_{i-1}^{-2}(e_i-\varepsilon_i)\|\Phi_{i-1}\|^2\}\\ &\qquad=o\left(\sum_{i=h+1}^{n}r_{i-1}^{-2}[\Phi'_{i-1}(\Theta_{i-1}-\Theta)]^2\right)+o\left(\sum_{i=h+1}^{n}r_{i-1}^{-2}(e_i-\varepsilon_i)^2\right)\\ &\qquad\quad+O(1)\quad\text{a.s.,}\end{aligned}\tag{4.7}$$

noting that $\|\Phi_t\|^2/r_t\leqq 1$. Moreover, $\sum_1^\infty r_t^{-2}\|\Phi_t\|^2<\infty$ since $r_t=r_0+\sum_1^t\|\Phi_i\|^2$. Hence

$$\sum_{i=h+1}^{n}r_{i-1}^{-2}\|\Phi_{i-1}\|^2\varepsilon_i^2=O\left(\sum_{i=h+1}^{n}r_{i-1}^{-2}\|\Phi_{i-1}\|^2\right)=O(1)\quad\text{a.s.,}\tag{4.8}$$

by Lemma 1(ii). By continuity and compactness, we can choose $\rho>0$ such that $\min_{|z|=1}\operatorname{Re}\{C(z)-(a+\rho)/2\}>0$. Moreover, by Lemma 2 (with $d=1$ and $\eta_t=\varepsilon_t$),

$$\Phi'_{i-1}(\Theta_{i-1}-\Theta)+(a+\rho)(e_i-\varepsilon_i)/2=-\{C(q^{-1})-\tfrac{1}{2}(a+\rho)\}(e_i-\varepsilon_i).\tag{4.9}$$

Since $\|\Phi_t\|^2/r_t\leqq 1$, it then follows from Lemma 3(ii) that

$$\begin{aligned}&\sum_{i=h+1}^{n}\{r_{i-1}^{-1}(e_i-\varepsilon_i)\Phi'_{i-1}(\Theta_{i-1}-\Theta)+(a/2)r_{i-1}^{-2}(e_i-\varepsilon_i)^2\|\Phi_{i-1}\|^2\}\\ &\qquad\leqq\sum_{i=h+1}^{n}r_{i-1}^{-1}(e_i-\varepsilon_i)\{\Phi'_{i-1}(\Theta_{i-1}-\Theta)+(a+\rho)(e_i-\varepsilon_i)/2\}\\ &\qquad\quad-(\rho/2)\sum_{i=h+1}^{n}r_{i-1}^{-1}(e_i-\varepsilon_i)^2\\ &\qquad\leqq-(\rho/2)\sum_{i=h+1}^{n}r_{i-1}^{-1}(e_i-\varepsilon_i)^2+O(1).\end{aligned}\tag{4.10}$$

Since $\Phi'_{i-1}(\Theta_{i-1}-\Theta)=-C(q^{-1})(e_i-\varepsilon_i)$,

$$\sum_{i=h+1}^{n}r_{i-1}^{-2}[\Phi'_{i-1}(\Theta_{i-1}-\Theta)]^2=O\left(\sum_{i=1}^{n}r_{i-1}^{-2}(e_i-\varepsilon_i)^2\right).\tag{4.11}$$

From (4.6)-(4.11), it follows that $Q_n\leqq\{-a\rho+o(1)\}\sum_{i=1}^n r_{i-1}^{-1}(e_i-\varepsilon_i)^2+O(1)$ almost surely, giving the desired conclusions (4.1) and (4.2). To prove (4.3), note that by (1.29a),

$$\begin{aligned}\sum_{i=1}^{n}\|\Theta_i-\Theta_{i-1}\|^2&=\sum_{i=1}^{n}a^2r_{i-1}^{-2}\|\Phi_{i-1}\|^2\{\varepsilon_i^2+2\varepsilon_i(e_i-\varepsilon_i)+(e_i-\varepsilon_i)^2\}\\ &\leqq(a^2+o(1))\sum_{i=1}^{n}r_{i-1}^{-1}(e_i-\varepsilon_i)^2+O(1)\quad\text{a.s.,}\end{aligned}$$

by (4.7) and (4.8), noting that $\|\Phi_t\|^2/r_t\leqq 1$. Hence (4.3) follows from (4.2).

(ii) Suppose that $\Theta\in D_i$ for all $i\geqq m(>h)$. For $i\geqq m$, since D_i is convex and $\Theta\in D_i$, it follows from (4.4) that

$$\begin{aligned}\|\Theta_i-\Theta_{i-1}\|^2&\leqq\|\Theta_{i-1}+ar_{i-1}^{-1}e_i\Phi_{i-1}-\Theta\|^2\\ &=\|\Theta_{i-1}-\Theta\|^2+2ar_{i-1}^{-1}e_i\Phi'_{i-1}(\Theta_{i-1}-\Theta)+a^2r_{i-1}^{-2}e_i^2\|\Phi_{i-1}\|^2,\end{aligned}$$

and therefore the same argument as before proves that (4.1) and (4.2) still hold. Suppose that $D_{i-1}\subset D_i$ for all $i\geqq m$. Then for $i\geqq m$, since $\Theta_{i-1}\in D_{i-1}\subset D_i$ and D_i is convex, it follows from (4.4) that $\|\Theta_i-\Theta_{i-1}\|^2\leqq\|(\Theta_{i-1}+ar_{i-1}^{-1}e_i\Phi_{i-1})-\Theta_i\|^2=a^2r_{i-1}^{-2}\|\Phi_{i-1}\|^2e_i^2$, and the same argument as before shows that (4.3) still holds. □

The ARMAX model (1.1) can be written as a linear regression model (1.12), and (1.29) represents the stochastic gradient algorithm to estimate the parameter vector Θ of this regression model. Making use of (4.1)-(4.3), Fuchs [16], [17] established the global convergence property (1.21) for the adaptive d-step ahead predictors defined by the explicit approach (1.18) in which the estimated polynomials are given by the stochastic gradient algorithm Θ_n, under assumptions (1.20), (1.25), and (1.26). An alternative approach for d-step ahead prediction is to express (1.1) as the regression model (1.9), in which $E(\eta_{n+d}|\mathscr{F}_n)=0$ while ψ_n is $\mathscr{F}_n$-measurable. The stochastic gradient algorithm estimating the parameter θ in the regression model $y_n=\theta'\psi_{n-d}+\eta_n$ (i.e., (1.9)) takes the form

$$\theta_n=\theta_{n-1}+(a/r_{n-d})\phi_{n-d}(y_n-\hat{y}_n), \tag{4.12a}$$

$$\phi_n=(y_n,\cdots,y_{n-p(d)+1},u_{n-\Delta+d},\cdots,u_{n-\Delta-k+2},\hat{y}_{n+d-1},\cdots,\hat{y}_{n+d-h})', \tag{4.12b}$$

$$r_n=r_{n-1}+\|\phi_n\|^2, \tag{4.12c}$$

$$\hat{y}_{n+d}=\theta_n'\phi_n, \tag{4.12d}$$

where $p(d)=p\vee(h-d+1)$. The extended stochastic Lyapunov function argument used in Corollary 1 can be modified to prove the global convergence of the adaptive predictors (4.12d) based on the stochastic gradient algorithm that does not involve interlacing in the implicit approach. This is the content of the following.

THEOREM 2. *Suppose that* $\min_{|z|=1}\operatorname{Re}\{C(z)-(d-1/2)a\}>0$ *and that the random disturbances* ε_n *in the linear stochastic system* (1.1) *satisfy* (1.3). *Consider the stochastic gradient algorithm* θ_n *defined by* (4.12) *for the implicit model* (1.9) *in which the* $\tilde{y}_i$ *and* η_i *are defined by* (1.6) *and* (1.7). *Then*

$$\limsup_{n\to\infty}\|\theta_n\|<\infty \quad a.s., \qquad \sum_{n=d+1}^{\infty}\|\theta_n-\theta_{n-1}\|^2<\infty \quad a.s., \tag{4.13}$$

$$\sum_{n=d+1}^{\infty}(\tilde{y}_n-\hat{y}_n)^2/r_{n-d}<\infty \quad a.s. \tag{4.14}$$

Consequently,

$$n^{-1}\sum_{t=d+1}^{n}(\tilde{y}_t-\hat{y}_t)^2\to 0 \quad a.s.\ on\ \left\{\limsup_{n\to\infty}n^{-1}\sum_{i=1}^{n}(y_i^2+u_i^2)<\infty\right\}. \tag{4.15}$$

If $B(z)\neq 0$ *and is stable, then on*

$$\{\limsup_{n\to\infty}n^{-1}\textstyle\sum_{t=1}^{n}y_t^2<\infty\}\cup\{\limsup_{n\to\infty}n^{-1}\textstyle\sum_{t=1}^{n}\hat{y}_t^2<\infty\},$$

$$n^{-1}\sum_{t=d+1}^{n}(\tilde{y}_t-\hat{y}_t)^2\to 0 \quad and \quad \sum_{t=1}^{n}(y_t^2+u_t^2+\hat{y}_t^2)=O(n) \quad a.s. \tag{4.16}$$

If $A(z)$ *is stable, then* (4.16) *holds on* $\{\limsup_{n\to\infty}n^{-1}\sum_{t=1}^{n}u_t^2<\infty\}$.

Proof. Let $Q_n=\|\theta_n-\theta\|^2$, $e_n=y_n-\hat{y}_n$. As in (4.6), it follows from (4.12a) that for $n>m$,

$$\begin{aligned}Q_n=Q_m&+2a\sum_{m+1}^{n}\{r_{i-d}^{-1}(e_i-\eta_i)\phi_{i-d}'(\theta_{i-1}-\theta)+(a/2)r_{i-d}^{-2}(e_i-\eta_i)^2\|\phi_{i-d}\|^2\}\\&+a^2\sum_{m+1}^{n}r_{i-d}^{-2}\|\phi_{i-d}\|^2\eta_i^2+2a\sum_{m+1}^{n}\eta_i\\&\times\{r_{i-d}^{-1}\phi_{i-d}'(\theta_{i-1}-\theta)+ar_{i-d}^{-2}(e_i-\eta_i)\|\phi_{i-d}\|^2\}.\end{aligned} \tag{4.17}$$

Note that r_{i-d}, ϕ_{i-d}, θ_{i-d}, and $e_i-\eta_i=\theta'\psi_{i-d}-\theta'_{i-d}\phi_{i-d}$ are $\mathscr{F}_{i-d}$-measurable. Since $\eta_i=\varepsilon_i+f_1\varepsilon_{i-1}+\cdots+f_{d-1}\varepsilon_{i-d+1}$, it follows from Lemma 1 that as $n\to\infty$,

$$\begin{aligned}&\sum_{m+1}^{n}\eta_i\{r_{i-d}^{-1}\phi'_{i-d}(\theta_{i-d}-\theta)+ar_{i-d}^{-2}(e_i-\eta_i)\|\phi_{i-d}\|^2\}\\&\qquad=o\left(\sum_{m+1}^{n}r_{i-d}^{-2}[\phi'_{i-d}(\theta_{i-d}-\theta)]^2\right)+o\left(\sum_{m+1}^{n}r_{i-d}^{-2}(e_i-\eta_i)^2\right)+O(1)\quad\text{a.s.,}\end{aligned}\tag{4.18}$$

$$\sum_{m+1}^{n}r_{i-d}^{-2}\|\phi_{i-d}\|^2\eta_i^2=O\left(\sum_{m+1}^{n}r_{i-d}^{-2}\|\phi_{i-d}\|^2\right)=O(1)\quad\text{a.s.,}\tag{4.19}$$

analogous to (4.7) and (4.8). Choosing $\rho>0$ such that $\operatorname{Re}\{C(z)-(d-\frac{1}{2})a-\rho\}>0$ for all $|z|=1$, we obtain by Lemma 2 that as in (4.9) and (4.10),

$$\begin{aligned}&\sum_{m+1}^{n}\{r_{i-d}^{-1}(e_i-\eta_i)\phi'_{i-d}(\theta_{i-d}-\theta)+(a/2)r_{i-d}^{-2}(e_i-\eta_i)^2\|\phi_{i-d}\|^2\}\\&\qquad\leqq-\sum_{m+1}^{n}r_{i-d}^{-1}(e_i-\eta_i)[\{C(q^{-1})-a/2-\rho\}(e_i-\eta_i)]-\rho\sum_{m+1}^{n}r_{i-d}^{-1}(e_i-\eta_i)^2.\end{aligned}\tag{4.20}$$

Noting that (4.17) involves $\theta_{i-1}-\theta$ instead of $\theta_{i-d}-\theta$, we write in the case $d>1$

$$\theta_{i-1}-\theta=(\theta_{i-d}-\theta)+\sum_{s=1}^{d-1}(\theta_{i-s}-\theta_{i-s-1}).\tag{4.21}$$

Fix $s=1,\cdots,d-1$. From (4.12a) and the inequality $|x'y|\leqq(\|x\|^2+\|y\|^2)/2$, it follows that

$$\begin{aligned}&\left|\sum_{m+1}^{n}r_{i-d}^{-1}(e_i-\eta_i)\phi'_{i-d}(\theta_{i-s}-\theta_{i-s-1})\right|\\&\qquad=\left|a\sum_{m+1}^{n}[r_{i-d}^{-1}(e_i-\eta_i)\phi_{i-d}]'[r_{i-s-d}^{-1}e_{i-s}\phi_{i-s-d}]\right|\\&\qquad\leqq(a/2)\left\{\sum_{m+1}^{n}r_{i-d}^{-2}(e_i-\eta_i)^2\|\phi_{i-d}\|^2+\sum_{m+1}^{n}r_{i-s-d}^{-2}e_{i-s}^2\|\phi_{i-s-d}\|^2\right\}\\&\qquad=(a+o(1))\sum_{m+1}^{n}r_{i-d}^{-2}(e_i-\eta_i)^2\|\phi_{i-d}\|^2+O(1)\quad\text{a.s.,}\end{aligned}\tag{4.22}$$

where the last relation above follows from an application of Lemma 1 to

$$\begin{aligned}\sum_{m+1}^{n}r_{i-s-d}^{-2}e_{i-s}^2\|\phi_{i-s-d}\|^2=&\sum_{m+1}^{n}r_{i-s-d}^{-2}(e_{i-s}-\eta_{i-s})^2\|\phi_{i-s-d}\|^2\\&+\sum_{m+1}^{n}r_{i-s-d}^{-2}\|\phi_{i-s-d}\|^2\eta_{i-s}^2\\&+2\sum_{m+1}^{n}r_{i-s-d}^{-2}\|\phi_{i-s-d}\|^2(e_{i-s}-\eta_{i-s})\eta_{i-s},\end{aligned}$$

noting that $e_{i-s}-\eta_{i-s}$ if $\mathscr{F}_{i-s-d}$-measurable and that $\|\phi_t\|^2/r_t\leqq 1$ and $\sum_1^\infty|\phi_t\|^2/r_t^2<\infty$.

Moreover, analogous to (4.22), we have

$$\begin{aligned}
&\left|\sum_{m+1}^{n} \eta_i r_{i-d}^{-1} \phi_{i-d}'(\theta_{i-s}-\theta_{i-s-1})\right| \\
&\qquad \leqq a\left|\sum_{m+1}^{n} \eta_i r_{i-d}^{-1} r_{i-s-d}^{-1} \phi_{i-d}' \phi_{i-s-d}(e_{i-s}-\eta_{i-s})\right| \\
&\qquad\quad + a\left|\sum_{m+1}^{n} [\eta_i r_{i-d}^{-1}\phi_{i-d}]'[\eta_{i-s} r_{i-s-d}^{-1}\phi_{i-s-d}]\right| \\
&\qquad = o\left(\sum_{m+1}^{n} r_{i-s-d}^{-1}(e_{i-s}-\eta_{i-s})^2\right) + O(1) + O\left(\sum_{m+1}^{n} r_{i-d}^{-2}\|\phi_{i-d}\|^2\eta_i^2\right) \\
&\qquad = o\left(\sum_{m+1}^{n} r_{i-d}^{-1}(e_i-\eta_i)^2\right) + O(1) \quad \text{a.s. by Lemma 1.}
\end{aligned} \tag{4.23}$$

Combining (4.21) with (4.20) and (4.22) gives

$$\begin{aligned}
&\sum_{m+1}^{n} \{r_{i-d}^{-1}(e_i-\eta_i)\phi_{i-d}'(\theta_{i-1}-\theta) + (a/2) r_{i-d}^{-2}(e_i-\eta_i)^2\|\phi_{i-d}\|^2\} \\
&\quad \leqq \{(d-1)a+o(1)\}\sum_{m+1}^{n} r_{i-d}^{-2}\|\phi_{i-d}\|^2(e_i-\eta_i)^2 + O(1) - \rho\sum_{m+1}^{n} r_{i-d}^{-1}(e_i-\eta_i)^2 \\
&\qquad - \sum_{m+1}^{n} r_{i-d}^{-1}(e_i-\eta_i)[\{C(q^{-1})-a/2-\rho\}(e_i-\eta_i)] \\
&\quad \leqq -(\rho+o(1))\sum_{m+1}^{n} r_{i-d}^{-1}(e_i-\eta_i)^2 + O(1) \\
&\qquad - \sum_{m+1}^{n} r_{i-d}^{-1}(e_i-\eta_i)[\{C(q^{-1})-(d-\tfrac{1}{2})a-\rho\}(e_i-\eta_i)] \\
&\quad \leqq -(\rho+o(1))\sum_{m+1}^{n} r_{i-d}^{-1}(e_i-\eta_i)^2 + O(1),
\end{aligned} \tag{4.24}$$

where the last inequality follows from Lemma 3(ii). As in (4.11), it follows from Lemma 2 that

$$\sum_{m+1}^{n} r_{i-d}^{-2}[\phi_{i-d}'(\theta_{i-d}-\theta)]^2 = O\left(\sum_{m+1}^{n} r_{i-d}^{-1}(e_i-\eta_i)^2\right). \tag{4.25}$$

From (4.17)-(4.19) together with (4.21) and (4.23)-(4.25), it follows that $Q_n + (2a\rho + o(1))\sum_{m+1}^{n} r_{i-d}^{-1}(e_i-\eta_i)^2 = O(1)$ almost surely, implying (4.13) and (4.14) as in the proof of Corollary 1. Note in this connection that $e_i - \eta_i = y_i - \hat{y}_i - \eta_i = \tilde{y}_i - \hat{y}_i$.

By (4.14) and the Kronecker lemma,

$$\begin{aligned}
\sum_{i=1}^{n} (y_i-\hat{y}_i-\eta_i)^2 &= O(1) \quad \text{on } \left\{\sup_n r_n < \infty\right\}, \\
&= o\left(\sum_{i=1}^{n-d} y_i^2 + \sum_{i=1}^{n-\Delta} u_i^2 + \sum_{i=1}^{n-1} \hat{y}_i^2\right) \quad \text{on } \left\{\sup_n r_n = \infty\right\}.
\end{aligned} \tag{4.26}$$

From Lemma 1(ii), (4.26), and the inequality

$$\sum_{i=1}^{n} \hat{y}_i^2 \leqq 2\sum_{i=1}^{n} (\hat{y}_i+\eta_i-y_i)^2 + 4\sum_{i=1}^{n} y_i^2 + 4\sum_{i=1}^{n} \eta_i^2, \tag{4.27}$$

it follows that

$$\sum_{i=1}^{n} \hat{y}_i^2 = O(n) \quad \text{a.s. on} \left\{ \limsup_{n\to\infty} n^{-1} \sum_{i=1}^{n} (y_i^2 + u_i^2) < \infty \right\}. \tag{4.28}$$

From (4.26) and (4.28), (4.15) follows.

Suppose that $B(z) \neq 0$ and is stable. by changing Δ if necessary, we can assume that $b_1 \neq 0$. Then by (1.1) and Lemma 5(i),

$$\sum_{i=1}^{n-\Delta} u_i^2 = O\left(\sum_{i=1}^{n} y_i^2 + \sum_{i=1}^{n} \varepsilon_i^2 \right) = O\left(\sum_{i=1}^{n} y_i^2 \right) + O(n) \quad \text{a.s.,} \tag{4.29}$$

and therefore

$$\sum_{i=1}^{n} u_i^2 = O(n) \quad \text{a.s. on} \left\{ \limsup_{n\to\infty} n^{-1} \sum_{i=1}^{n} y_i^2 < \infty \right\}. \tag{4.30}$$

From (4.26) and (4.29), it follows that

$$\begin{aligned} \sum_{i=1}^{n} y_i^2 &\leqq 2 \sum_{i=1}^{n} (y_i - \hat{y}_i - \eta_i)^2 + 4 \sum_{i=1}^{n} \hat{y}_i^2 + 4 \sum_{i=1}^{n} \eta_i^2 \\ &= o\left(\sum_{i=1}^{n-d} y_i^2 \right) + O\left(\sum_{i=1}^{n} \hat{y}_i^2 \right) + O(n) \quad \text{a.s.,} \end{aligned}$$

and therefore

$$\sum_{i=1}^{n} y_i^2 = O(n) \quad \text{a.s. on} \left\{ \limsup_{n\to\infty} n^{-1} \sum_{i=1}^{n} \hat{y}_i^2 < \infty \right\}. \tag{4.31}$$

On the event $\{\limsup_{n\to\infty} n^{-1} \sum_{i=1}^{n} y_i^2 < \infty\} \cup \{\limsup_{n\to\infty} n^{-1} \sum_{i=1}^{n} \hat{y}_i^2 < \infty\}$, it follows from (4.28), (4.30), and (4.31) that $\sum_{i=1}^{n} (y_i^2 + u_i^2 + \hat{y}_i^2) = O(n)$ almost surely, and therefore $\sum_{i=1}^{n} (y_i - \hat{y}_i - \eta_i)^2 = o(n)$ almost surely by (4.26).

Now assume that $A(z)$ is stable. Then by (1.1) and Lemma 5(i),

$$\sum_{i=1}^{n} y_i^2 = O\left(\sum_{i=1}^{n-\Delta} u_i^2 \right) + O\left(\sum_{i=1}^{n} \varepsilon_i^2 \right) = O\left(\sum_{i=1}^{n-\Delta} u_i^2 \right) + O(n) \quad \text{a.s.}$$

Hence on the event $\{\limsup_{n\to\infty} n^{-1} \sum_{i=1}^{n} u_i^2 < \infty\}$, $\sum_{i=1}^{n} y_i^2 = O(n)$ almost surely, and therefore $\sum_{i=1}^{n} \hat{y}_i^2 = O(n)$ almost surely by (4.28) and $\sum_{i=1}^{n} (y_i - \hat{y}_i - \eta_i)^2 = o(n)$ almost surely by (4.26). □

Noting that although the stochastic gradient algorithm leads to the adaptive controller (1.23)-(1.24) that is self-optimizing, its rate of convergence has been found in numerical studies to be much slower than that of the more commonly used "least squares iterations," Sin and Goodwin [21] suggest combining the ideas underlying both algorithms into what they call a "modified least squares" algorithm for the case $d = \Delta = 1$. Their algorithm has been modified and extended to general d by Zhang [13] who, defining $\bar{\phi}_n$ as in (3.8c), introduces the recursions

$$\bar{\theta}_n = \bar{\theta}_{n-d} + \alpha_{n-d} P_{n-d} \bar{\phi}_{n-d} (y_n - \bar{\theta}'_{n-d} \bar{\phi}_{n-d}), \tag{4.32a}$$

$$r_n = r_{n-1} + \|\bar{\phi}_n\|^2, \qquad R_n = (P_{n-d}^{-1} + \bar{\phi}_n \bar{\phi}'_n)^{-1}, \tag{4.32b}$$

$$\begin{aligned} &P_n = R_n \quad \text{and } \alpha_n = 1, \quad \text{if } r_n \operatorname{tr}(R_n) \leqq K_1 \quad \text{and } \bar{\phi}'_n P_{n-d} \bar{\phi}_n \leqq K_2, \\ &P_n = (r_{n-d}/r_n) P_{n-d} \quad \text{and } \alpha_n = 1/(1 + \bar{\phi}'_n P_n \bar{\phi}_n), \quad \text{otherwise,} \end{aligned} \tag{4.32c}$$

where K_1, K_2 are prescribed constants. Under assumptions (1.20) and (3.10), it is shown in [13] that the adaptive predictors $\hat{y}_{n+d} = \bar{\theta}'_n \bar{\phi}_n$ associated with (4.32) satisfy the global convergence property (1.21). The modified least squares algorithm of Sin and Goodwin [21] for the case $d = \Delta = 1$ is similar, but does not include the condition $\bar{\phi}'_n P_{n-1} \bar{\phi}_n \leqq K_2$ as in (4.32c) because Solo's [8] result (A6) has been used to conclude that $\bar{\phi}'_n P_{n-1} \bar{\phi}_n \to 0$. Zhang [13] finds (A6) questionable and further modifies the Sin-Goodwin algorithm to patch this gap and to extend to general d.

Instead of using matrix gain (3.8b) of the extended least squares algorithm, which we have studied in § 3, Zhang's modified least squares algorithm changes this gain by not including $\bar{\phi}_n \bar{\phi}'_n$ into the sum and deflating the matrix by a scalar multiple whenever certain conditions are not met. Note that the scalar gain $1/r_n = 1/(r_0 + \sum_1^n \|\bar{\phi}_i\|^2)$ of the stochastic gradient algorithm also appears in the condition $r_n \operatorname{tr}(R_n) \leq K_1$ and in the dampening factor r_{n-d}/r_n of (4.32c). Our results in Theorems 1 and 2 suggest a simpler and more direct way of combining the stochastic gradient algorithm with the extended least squares algorithm to produce a globally convergent adaptive predictor. This is the content of the following result.

COROLLARY 2. *Suppose that the random disturbances ε_n in the linear system* (1.1) *satisfy* (1.3) *and that the outputs y_n and inputs u_n satisfy* (1.20). *Assume that* (3.10) *holds and that* $\min_{|z|=1} \operatorname{Re}\{C(z) - (d - \frac{1}{2})a\} > 0$ *for some $a > 0$. Define the stochastic gradient algorithm θ_n by* (4.12) *and the extended least squares algorithm $\bar{\theta}_n$ by* (3.8). *Let $\hat{y}_{n+d}$ be the adaptive d-step ahead predictor associated with the stochastic gradient algorithm, as in* (4.12d). *Take $0 < \delta < 1$ and define*

$$\hat{y}^*_{n+d} = \bar{\theta}'_n \bar{\phi}_n \quad \text{if } \bar{\phi}'_n P_n \bar{\phi}_n \leqq \delta, \qquad = \hat{y}_{n+d} \quad \text{if } \bar{\phi}'_n P_n \bar{\phi}_n > \delta, \tag{4.33}$$

where P_n and $\bar{\phi}_n$ are given in (3.8b) *and* (3.8c). *Then*

$$n^{-1} \sum_{i=1}^n (\tilde{y}_{i+d} - \hat{y}^*_{i+d})^2 \to 0 \quad a.s., \tag{4.34}$$

$$\sum_{i=1}^n I_{\{\bar{\phi}'_i P_i \bar{\phi}_i > \delta\}} = O(\log n) \quad a.s. \tag{4.35}$$

Proof. By Theorem 1, $\sum_1^n \|\bar{\phi}_i\|^2 = O(n)$ almost surely since (1.20) holds, and

$$\sum_{i=1}^n (\tilde{y}_{i+d} - \bar{\theta}'_i \bar{\phi}_i)^2 I_{\{\bar{\phi}'_i P_i \bar{\phi}_i \leqq \delta\}} = O(\log \lambda_{\max}(P_n^{-1})) = O(\log n) \quad \text{a.s.} \tag{4.36}$$

By (4.15) and (1.20),

$$\sum_{i=1}^n (\tilde{y}_{i+d} - \hat{y}_{i+d})^2 = o(n) \quad \text{a.s.} \tag{4.37}$$

From (4.33), (4.36), and (4.37), (4.34) follows.

To prove (4.35), note that for $\nu = 0, \cdots, d-1$,

$$\sum_{i \leqq n, i \equiv \nu (\operatorname{mod} d)} \bar{\phi}'_i (P_i^{-1})^{-1} \bar{\phi}_i = O\left(\log\left(2 + \sum_{i=1}^n \|\bar{\phi}_i\|^2\right)\right) \quad \text{a.s.}, \tag{4.38}$$

by (3.8b) and Lemma 4(i). Since $\sum_1^n \|\bar{\phi}_i\|^2 = O(n)$ almost surely, (4.35) follows from (4.38). □

In view of (4.35), the stochastic gradient component of the adaptive predictor (4.33) is used very infrequently, only at a relative frequency of $O(n^{-1} \log n)$ within n

stages. At other times, the extended least squares component of (4.33) is used and the cumulative squared difference between the adaptive predictor and the optimal predictor at these times is of the order $O(\log n)$, as in (4.36). We can introduce at these times another modification to keep $\sum_{i=1}^{n} (\tilde{y}_{i+d} - \hat{y}_{i+d})^2$ within $O((\log n)^3)$ almost surely if the inputs and outputs eventually have finite moment generating functions, i.e., if $\limsup_{t\to\infty} E \exp(\lambda|u_t|) < \infty$ and $\limsup_{t\to\infty} E \exp(\lambda|y_t|) < \infty$ for some $\lambda > 0$. This is the content of the following corollary.

COROLLARY 3. *Suppose that the random disturbances* ε_n *in* (1.1) *satisfy* (1.3) *and that the outputs* y_n *and inputs* u_n *satisfy* (1.20) *and*

$$\limsup_{i\to\infty} \{E \exp(\lambda|y_i|^\alpha) + E \exp(\lambda|u_i|^\alpha)\} < \infty \quad \text{for some } \lambda > 0 \text{ and } \alpha > 0. \tag{4.39}$$

Assume that (3.10) *holds. Define the extended least squares algorithm* $\bar{\theta}_n$ *by* (3.8). *Take* $0 < \delta < 1$ *and define*

$$\begin{aligned} \hat{y}_{n+d} &= \bar{\theta}'_n \bar{\phi}_n \quad \text{if } \bar{\phi}'_n P_n \bar{\phi}_n \leqq \delta, \\ &= (\log n)^{1/\alpha} \quad \text{if } \bar{\phi}'_n P_n \bar{\phi}_n > \delta, \end{aligned} \tag{4.40}$$

where P_n *and* $\bar{\phi}_n$ *are given in* (3.8b) *and* (3.8c). *Then* (4.35) *holds and*

$$\sum_{i=1}^{n} (\tilde{y}_{i+d} - \hat{y}_{i+d})^2 = O((\log n))^{1+2/\alpha}) \quad \text{a.s.} \tag{4.41}$$

Proof. First note from the proof of Corollary 2 that (4.36), (4.38), and (4.35) still hold in the present case. By (3.10), $\bar{C}(z)(=\bar{F}(z)C(z))$ is stable and therefore $C(z)$ is stable. From (4.39) it follows that

$$\sum_{n=1}^{\infty} (P\{|u_n| > 2(\lambda^{-1}\log n)^{1/\alpha}\} + P\{|y_n| > 2(\lambda^{-1}\log n)^{1/\alpha}\}) = \sum_{n=1}^{\infty} O(\exp(-2\log n)) < \infty,$$

and therefore $\max(|u_n|, |y_n|) = O((\log n)^{1/\alpha})$ almost surely by the Borel-Cantelli lemma. Since $C(z)$ is stable, it then follows from (1.6) and Lemma 5(i) that $\tilde{y}_{n+d} = O((\log n)^{1/\alpha})$ almost surely. Hence by (4.40),

$$\sum_{i=1}^{n} (\tilde{y}_{i+d} - \hat{y}_{i+d})^2 I_{\{\bar{\phi}'_i P_i \bar{\phi}_i > \delta\}} = O\left((\log n)^{2/\alpha} \sum_{i=1}^{n} I_{\{\bar{\phi}'_i P_i \bar{\phi}_i > \delta\}}\right) \quad \text{a.s.} \tag{4.42}$$

From (4.35), (4.36), and (4.42), (4.41) follows. □

5. Consistent parameter estimation and monitored recursive maximum likelihood. In this section we will assume that $C(z)$ is stable. Suppose that we are able to find strongly consistent estimators Θ_n of the parameter vector Θ defined in (1.11). Then under assumption (1.20), the adaptive d-step ahead predictor $\hat{y}_{n+d}$ constructed from Θ_n by the explicit approach (1.18) satisfies the global convergence property (1.21). Here we will make use of the consistent estimators Θ_n in another way to provide adaptive predictors $\hat{y}_{n+d}$ that satisfy

$$\begin{aligned} &\limsup_{n\to\infty} \frac{\sum_{i=1}^{n} (\tilde{y}_{i+d} - \hat{y}_{i+d})^2}{\log n} \\ &\quad \leqq (2d-1)(p(d)+k+d-1+h) \limsup_{i\to\infty} E(\eta_i^2 | \mathcal{F}_{i-d}) \quad \text{a.s.,} \end{aligned} \tag{5.1}$$

where $p(d) = p \vee (h-d+1)$ and η_i is defined in (1.7). Such adaptive predictors are constructed by the implicit approach using the monitored recursive maximum likelihood algorithm defined below. Note that (5.1) is a stronger conclusion than (3.14) on adaptive predictors based on the extended least squares algorithm. Moreover, while the extended least squares algorithm needs the positive real assumption (3.10) for

(3.14) to hold, the monitored recursive maximum likelihood algorithm does not need such an assumption for (5.1).

The consistent estimators Θ_n will be used to provide "confidence sets" S_n for the parameter vector θ, defined in (1.8), of the implicit system such that S_n shrinks to θ as $n \to \infty$. For example, consider the AML algorithm Θ_n defined in (1.13) with pseudoregression vectors Φ_n. Under assumptions (1.15) and (1.17), it follows from Theorem 1 of [9] that $\{\log(\sum_1^n \|\Phi_i\|^2)\}/\lambda_{\min}(\sum_1^n \Phi_i\Phi_i') \to 0$ almost surely and that

$$\Theta_n - \Theta = O\left(\left\{\log\left(\sum_1^n \|\Phi_i\|^2\right)\right\}^{1/2} \Big/ \lambda_{\min}^{1/2}\left(\sum_1^n \Phi_i\Phi_i'\right)\right) \quad \text{a.s.} \tag{5.2}$$

Since the components of θ can be expressed as smooth functions of those of Θ, Θ_n induces a strongly consistent estimator θ_n^* of θ; in fact, (5.2) implies that

$$\theta_n^* - \theta = O\left(\left\{\log\left(\sum_{i=1}^n \|\Phi_i\|^2\right)\right\}^{1/2} \Big/ \lambda_{\min}^{1/2}\left(\sum_{i=1}^n \Phi_i\Phi_i'\right)\right) \quad \text{a.s.} \tag{5.3}$$

Hence we can define S_n to be a cube with center θ_n^* and width $\lambda_{\min}^{-1/3}(I + \sum_1^n \Phi_i\Phi_i')$. Then by (5.3),

$$P\{\theta \in S_n \text{ for all large } n\} = 1, \tag{5.4}$$

and as $n \to \infty$ the width of S_n converges to 0 almost surely.

The consistent estimators θ_n^* and the associated confidence sets S_n need only be updated occasionally at times $n(1) < n(2) < \cdots$ for monitoring the recursive maximum likelihood algorithm θ_n that we now introduce. The basic ideas underlying the algorithm θ_n are (i) to extend the RML2 algorithm (1.33) to the implicit system (1.9), and (ii) to constrain (monitor) the algorithm so that it lies inside $S_{n(j)}$ for $n(j) \leqq n < n(j+1)$. The projection which we use to constrain θ_n is taken with respect to the norm induced by the positive definite matrix P_{n-d}^{-1} defined in (5.6d) below, instead of the usual Euclidean norm. For $x \in R^{p(d)+k+d-1+h}$ and $n(j) \leqq n < n(j+1)$, let $\pi_n(x)$ denote the unique solution of the quadratic programming problem

$$(\pi_n(x) - x)' P_{n-d}^{-1}(\pi_n(x) - x) = \min_{y \in S_{n(j)}} \{(y-x)' P_{n-d}^{-1}(y-x)\}, \tag{5.5}$$

i.e.,$\pi_n(x)$ is the projection of x into $S_{n(j)}$ with respect to the norm induced by P_{n-d}^{-1}. It is convenient to choose $S_{n(j)}$ to be a cube so that we have linear constraints for the quadratic programming problem (5.5), which can be handled by simple computational methods (cf. [22]). Define $\theta_n = (\hat{g}_{n,1}, \cdots, \hat{g}_{n,p(d)}, \hat{b}_{n,1}, (\widehat{fb})_{n,2}, \cdots, (\widehat{fb})_{n,k+d-1}, -\hat{c}_{n,1}, \cdots, -\hat{c}_{n,h})'$ for $n > n(1)$ by the recursion

$$\theta_n = \pi_n(\theta_{n-1} + P_{n-d}\xi_{n-d}(y_n - \hat{y}_n)), \tag{5.6a}$$

$$\begin{aligned} &\xi_n + \hat{c}_{n,1}\xi_{n-1} + \cdots + \tilde{c}_{n,h}\xi_{n-h} = \phi_n, \quad \text{where} \\ &\phi_n = (y_n, \cdots, y_{n-p(d)+1}, u_{n-\Delta+d}, \cdots, u_{n-k-\Delta+2}, \hat{y}_{n+d-1}, \cdots, \hat{y}_{n+d-h})', \end{aligned} \tag{5.6b}$$

$$\hat{y}_{n+d} = \theta_n' \phi_n, \tag{5.6c}$$

$$P_n^{-1} = P_{n-1}^{-1} + \xi_n \xi_n' + \rho_n I, \tag{5.6d}$$

where

$$\rho_n \geqq 0 \quad \text{is } \mathscr{F}_n\text{-measurable with } \sup_n \rho_n < \infty \quad \text{a.s.} \tag{5.6e}$$

The following theorem, which is analogous to Theorem 1 on the extended least squares algorithm, gives asymptotic properties of the adaptive predictors (5.6c) associated with the monitored recursive maximum likelihood algorithm. These results are used to establish the conclusion (5.1) for the choice $\rho_n = 1/n$ in (5.6d) under certain conditions on the input-output data and the stopping times $n(j)$ in Corollary 4 below.

THEOREM 3. *Suppose that $C(z)$ is stable and that the random disturbances ε_n in the linear stochastic system* (1.1) *satisfy assumption* (1.3). *Let $n(1) < n(2) < \cdots$ be stopping times with respect to $\{\mathscr{F}_t\}$ and let $S_{n(j)}$ be an $\mathscr{F}_{n(j)}$-measurable, closed and convex set such that*

$$P\{\theta \in S_{n(j)} \textit{ for all large } j\} = 1 \quad \textit{and} \quad \lim_{j\to\infty} (\textit{diameter of } S_{n(j)}) = 0 \quad \textit{a.s.} \tag{5.7}$$

Define the monitored recursive maximum likelihood algorithm θ_n by (5.6), *where π_n is given by* (5.5) *for $n(j) \leqq n < n(j+1)$, and define $\tilde{y}_t$ and η_t by* (1.6) *and* (1.7).

(i) *Suppose that* $\sup_n |\varepsilon_n| < \infty$ *almost surely. Then on the event* $\{\lambda_{\max}(P_n^{-1}) \to \infty$ *and* $\xi'_n P_n \xi_n \to 0\}$,

$$\begin{aligned}\sum_{i\leqq n} (\tilde{y}_{i+d} - \hat{y}_{i+d})^2 \leqq (2d-1)&\left\{\limsup_{i\to\infty} E(\eta_i^2 \mid \mathscr{F}_{i-d}) + o(1)\right\} \log\det(P_n^{-1}) \\ &+ o\left(\sum_{i=1}^{n} \rho_i\right) \\ &+ o\left(\sum_{j:n(j)\leqq n+d} \sum_{s=1}^{h+d-1} [\|\xi_{n(j)-s}\|^2 + \|\xi_{n(j)-s}\|]\right) \quad \textit{a.s.}\end{aligned} \tag{5.8}$$

(ii) *Suppose that the $n(j)$ are stopping times with respect to $\{\mathscr{F}_{t-d+1}\}$ (i.e., $\{n(j)=t\} \in \mathscr{F}_{t-d+1}$) and such that $n(j+1) - n(j) \geqq d$. Then* (5.8) *still holds on* $\{\lambda_{\max}(P_n^{-1}) \to \infty$ *and* $\xi'_n P_n \xi_n \to 0\}$.

Proof. Let $Q_n = (\theta_n - \theta)' P_{n-d}^{-1} (\theta_n - \theta)$. For $n(j) \leqq n < n(j+1)$, since $\pi_n(x)$ is the projection of x into the closed convex set $S_{n(j)}$ with respect to the norm induced by P_{n-d}^{-1}, it follows from (5.6a) that if $\theta \in S_{n(j)}$,

$$\begin{aligned} Q_n &\leqq (\theta_{n-1} - \theta + P_{n-d}\xi_{n-d}(y_n - \hat{y}_n))' P_{n-d}^{-1} (\theta_{n-1} - \theta + P_{n-d}\xi_{n-d}(y_n - \hat{y}_n)) \\ &= (\theta_{n-1} - \theta)' P_{n-d}^{-1} (\theta_{n-1} - \theta) + 2\xi'_{n-d}(\theta_{n-1} - \theta)(y_n - \hat{y}_n) + \xi'_{n-d} P_{n-d} \xi_{n-d} (y_n - \hat{y}_n)^2 \\ &= Q_{n-1} + \rho_{n-d}(\theta_{n-1} - \theta)^2 + [\xi'_{n-d}(\theta_{n-1} - \theta)]^2 + 2\xi'_{n-d}(\theta_{n-1} - \theta)(y_n - \hat{y}_n) \\ &\quad + \xi'_{n-d} P_{n-d} \xi_{n-d} (y_n - \hat{y}_n)^2, \end{aligned} \tag{5.9}$$

noting that $P_t^{-1} = P_{t-1}^{-1} + \xi_t \xi'_t + \rho_t I$. Therefore on the event $E \triangleq \{\theta \notin S_{n(j)}$ for finitely many j's$\}$,

$$\begin{aligned} Q_n &\leqq \sum_{i=n(1)+1}^{n} [\xi'_{i-d}(\theta_{i-1} - \theta)]^2 + 2 \sum_{i=n(1)+1}^{n} \xi'_{i-d}(\theta_{i-1} - \theta)(y_i - \hat{y}_i) \\ &\quad + \sum_{i=n(1)+1}^{n} \xi'_{i-d} P_{i-d} \xi_{i-d} (y_i - \hat{y}_i)^2 + \sum_{i=n(1)+1}^{n} \rho_{i-d}(\theta_{i-1} - \theta)^2 + O(1) \\ &\leqq - \sum_{i=n(1)+1}^{n} [\xi'_{i-d}(\theta_{i-1} - \theta)]^2 + 2 \sum_{i=n(1)+1}^{n} \xi'_{i-d}(\theta_{i-1} - \theta) \\ &\quad \times [\xi'_{i-d}(\theta_{i-1} - \theta) + (y_i - \hat{y}_i - \eta_i)] \\ &\quad + 2 \sum_{i=n(1)+1}^{n} \xi'_{i-d}(\theta_{i-1} - \theta)\eta_i + \sum_{i=n(1)+1}^{n} \xi'_{i-d} P_{i-d} \xi_{i-d} (y_i - \hat{y}_i)^2 + o\left(\sum_{i=1}^{n-d} \rho_i\right) \\ &\quad + O(1) \quad \text{a.s.,} \end{aligned} \tag{5.10}$$

noting that $\theta_{i-1} - \theta \to 0$ almost surely by (5.7).

Let $C_t(q^{-1}) = 1 + \hat{c}_{t,1}q^{-1} + \cdots + \hat{c}_{t,h}q^{-h}$, $\hat{c}_{t,0} = 1$. From (2.3),

$$
\begin{aligned}
C(q^{-1})[\xi'_{i-d}(\theta_{i-1}-\theta)+(y_i-\hat{y}_i-\eta_i)] &= C(q^{-1})[\xi'_{i-d}(\theta_{i-1}-\theta)]-\phi'_{i-d}(\theta_{i-d}-\theta)\\
&= [C(q^{-1})-C_{i-d}(q^{-1})][\xi'_{i-d}(\theta_{i-1}-\theta)]\\
&\quad + \sum_{r=1}^{h} \hat{c}_{i-d,r}\xi'_{i-d-r}(\theta_{i-1-r}-\theta_{i-1})\\
&\quad + \phi'_{i-d}(\theta_{i-1}-\theta_{i-d}),
\end{aligned}
\tag{5.11}
$$

noting that $C_{i-d}(q^{-1})\xi'_{i-d}\theta = \phi'_{i-d}\theta$ and $(\sum_{r=0}^{h}\hat{c}_{i-d,r}\xi'_{i-d-r})\theta_{i-1} = \phi'_{i-d}\theta_{i-1}$ by (5.6b). Since $C(z)$ is stable and $\hat{c}_{i-d,r} \to c_r$ almost surely as $i \to \infty$ for $r = 1, \cdots, h$, and since $\theta_{i-1}-\theta_{i-s} = \sum_{j=i-s+1}^{i-1}(\theta_j-\theta_{j-1})$ for $s \geqq 1$, it follows from (5.11), (5.6b), and Lemma 5(i) that

$$
\begin{aligned}
&\sum_{i=n(1)+1}^{n}[\xi'_{i-d}(\theta_{i-1}-\theta)+(y_i-\hat{y}_i-\eta_i)]^2\\
&\quad = o\left(\sum_{i=n(1)+1}^{n}[\xi'_{i-d}(\theta_{i-1}-\theta)]^2\right)+O(1)\\
&\quad\quad + O\left(\sum_{i=1}^{n}\sum_{r=-d+1}^{h-1}[\xi'_{i-d-r}(\theta_i-\theta_{i-1})]^2\right) \quad \text{a.s.}
\end{aligned}
\tag{5.12}
$$

For $n(j) < i < n(j+1)$, $\theta_{i-1} \in S_{n(j)}$ and therefore $(\theta_i-\theta_{i-1})'P_{i-d}^{-1}(\theta_i-\theta_{i-1}) \leqslant \xi'_{i-d}P_{i-d}\xi_{i-d}(y_i-\hat{y}_i)^2$ by (5.6a). Hence by the Schwarz inequality, on $E_1 \triangleq E \cap \{\lim_{n\to\infty}\xi'_nP_n\xi_n = 0\}$,

$$
\begin{aligned}
&\sum_{i\notin\{n(1),n(2),\cdots\}}^{n}\ \sum_{r=-d+1}^{h-1}[\xi'_{i-d-r}P_{i-d}^{1/2}P_{i-d}^{-1/2}(\theta_i-\theta_{i-1})]^2\\
&\quad \leqq \sum_{i=n(1)+1}^{n}\left(\sum_{r=-d+1}^{h-1}\xi'_{i-d-r}P_{i-d}\xi_{i-d-r}\right)\xi'_{i-d}P_{i-d}\xi_{i-d}(y_i-\hat{y}_i)^2\\
&\quad = o\left(\sum_{i=n(1)+1}^{n}\xi'_{i-d}P_{i-d}\xi_{i-d}(y_i-\hat{y}_i)^2\right)+O(1)
\end{aligned}
\tag{5.13}
$$

by Lemma 4(ii). Hence by (5.12) and (5.13),

$$
\begin{aligned}
&\sum_{i=n(1)+1}^{n}[\xi'_{i-d}(\theta_{i-1}-\theta)+(y_i-\hat{y}_i-\eta_i)]^2\\
&\quad = o\left(\sum_{i=n(1)+1}^{n}[\xi'_{i-d}(\theta_{i-1}-\theta)]^2\right)+O(1)\\
&\quad\quad + o\left(\sum_{i=n(1)+1}^{n}\xi'_{i-d}P_{i-d}\xi_{i-d}(y_i-\hat{y}_i)^2\right)\\
&\quad\quad + o\left(\sum_{j:n(j)\leqq n}\ \sum_{r=-d+1}^{h-1}\|\xi_{n(j)-d-r}\|^2\right) \quad \text{a.s. on } E_1,
\end{aligned}
\tag{5.14}
$$

noting that $\theta_j-\theta_{j-1}\to 0$ almost surely in view of (5.7).

Since $(y_i-\hat{y}_i)^2=(y_i-\hat{y}_i-\eta_i)^2+2(y_i-\hat{y}_i-\eta_i)\eta_i+\eta_i^2$ and since $y_i-\hat{y}_i-\eta_i$ and $\xi'_{i-d}P_{i-d}\xi_{i-d}(\leqq 1)$ are $\mathscr{F}_{i-d}$-measurable, it follows from Lemma 1 that on E_1,

$$\begin{aligned}
&\sum_{i=n(1)+1}^{n} \xi'_{i-d}P_{i-d}\xi_{i-d}(y_i-\hat{y}_i)^2\\
&\quad\leqq \{\limsup_{i\to\infty} E(\eta_i^2\,|\,\mathscr{F}_{i-d})+o(1)\}\sum_{i=n(1)+1}^{n}\xi'_{i-d}P_{i-d}\xi_{i-d}\\
&\qquad+o\left(\sum_{i=n(1)+1}^{n}(y_i-\hat{y}_i-\eta_i)^2\right)+O(1) \qquad (5.15)\\
&\quad=\left\{\limsup_{t\to\infty} E(\eta_i^2\,|\,\mathscr{F}_{i-d})+o(1)\right\}\log\det P_{n-d}^{-1}\\
&\qquad+o\left(\sum_{i=n(1)+1}^{n}(y_i-\hat{y}_i-\eta_i)^2\right)+O(1)\quad\text{a.s.,}
\end{aligned}$$

by Lemma 4(i). Let $0<\lambda<\frac{1}{2}$. Using the inequality $A^2\leqq(1+\lambda^2)B^2+(1+\lambda^{-2})(A+B)^2$, we obtain from (5.14) and (5.15) that on E_1,

$$\begin{aligned}
&\sum_{i=n(1)+1}^{n}(y_i-\hat{y}_i-\eta_i)^2\\
&\quad\leqq(1+\lambda^2+o(1))\sum_{i=n(1)+1}^{n}[\xi'_{i-d}(\theta_{i-1}-\theta)]^2 \qquad (5.16)\\
&\qquad+o(\log\det P_{n-d}^{-1})+o\left(\sum_{j:n(j)\leqq n}\sum_{s=1}^{h+d-1}\|\xi_{n(j)-s}\|^2\right)+O(1)\quad\text{a.s.}
\end{aligned}$$

Moreover, using the inequality $AB\leqq(\lambda^2A^2+\lambda^{-2}B^2)/2$, we obtain from (5.14) and (5.15) that on E_1

$$\begin{aligned}
&\sum_{i=n(1)+1}^{n}[\xi'_{i-d}(\theta_{i-1}-\theta)][\xi'_{i-d}(\theta_{i-1}-\theta)+(y_i-\hat{y}_i-\eta_i)]\\
&\leqq(\lambda^2+o(1))\sum_{i=n(1)+1}^{n}[\xi'_{i-d}(\theta_{i-1}-\theta)]^2 \qquad (5.17)\\
&\qquad+o(\log\det P_{n-d}^{-1})+o\left(\sum_{j:n(j)\leqq n}\sum_{s=1}^{h+d-1}\|\xi_{n(j)-s}\|^2\right)+o\left(\sum_{i=1}^{n-d}\rho_i\right)+O(1)\quad\text{a.s.}
\end{aligned}$$

Writing $\theta_{i-1}-\theta=\theta_{i-d}-\theta+\sum_{r=1}^{d-1}(\theta_{i-r}-\theta_{i-r-1})$, we now proceed to show that on E_1,

$$\begin{aligned}
&2\left|\sum_{i=n(1)+1}^{n}\xi'_{i-d}(\theta_{i-1}-\theta)\eta_i\right|\\
&\quad\leqq 2(d-1)\left\{\limsup_{i\to\infty}E(\eta_i^2\,|\,\mathscr{F}_{i-d})+o(1)\right\}\log\det P_{n-d}^{-1}\\
&\qquad+o\left(\sum_{i=n(1)+1}^{n}[\xi'_{i-d}(\theta_{i-1}-\theta)]\right)+o\left(\sum_{j:n(j)\leqq n}\sum_{s=1}^{h+d-1}\|\xi_{n(j)-s}\|^2\right) \qquad (5.18)\\
&\qquad+o\left(\sum_{j:n(j)\leqq n}\sum_{r=1}^{d-1}\|\xi_{n(j)+r-d}\|\,|\eta_{n(j)+r}|\right)+O(1)\quad\text{a.s.}
\end{aligned}$$

Since $\xi'_{i-d}(\theta_{i-d}-\theta)$ is $\mathscr{F}_{i-d}$-measurable and since $\eta_i=\varepsilon_i+f_1\varepsilon_{i-1}+\cdots+f_{d-1}\varepsilon_{i-d+1}$, an application of Lemma 1(i) gives

$$\begin{aligned}\sum_{i=n(1)+1}^{n}\xi'_{i-d}(\theta_{i-d}-\theta)\eta_i&=o\left(\sum_{i=n(1)+1}^{n}[\xi'_{i-d}(\theta_{i-d}-\theta)]^2\right)+O(1)\\ &=o\left(\sum_{i=n(1)+1}^{n}[\xi'_{i-d}(\theta_{i-1}-\theta)]^2\right)\\ &\quad+o\left(\sum_{r=1}^{d-1}\sum_{i=n(1)+1}^{n}[\xi'_{i-d}(\theta_{i-r}-\theta_{i-r-1})]^2\right)\\ &\quad+O(1)\quad\text{a.s.}\end{aligned}\tag{5.19}$$

For fixed $r=1,\cdots,d-1$, we have analogous to (5.13) that on E_1

$$\begin{aligned}\sum_{i=n(1)+1}^{n}[\xi'_{i-d}(\theta_{i-r}-\theta_{i-r-1})]^2&=o\left(\sum_{i=n(1)+1}^{n-r}\xi'_{i-d}P_{i-d}\xi_{i-d}(y_i-\hat{y}_i)^2\right)\\ &\quad+o\left(\sum_{j:n(j)\leqq n-r}\|\xi_{n(j)+r-d}\|^2\right)+O(1).\end{aligned}\tag{5.20}$$

Moreover, since $\theta_{t-1}\in S_{n(j)}$ if $n(j)<t<n(j+1)$, we obtain by the Schwarz inequality and (5.6a) that

$$\begin{aligned}&\sum_{i\notin\{n(1)+r,n(2)+r,\cdots\}}^{n}|\xi'_{i-d}P^{1/2}_{i-r-d}P^{-1/2}_{i-r-d}(\theta_{i-r}-\theta_{i-r-1})\eta_i|\\ &\leqq\sum_{i\notin\{n(1)+r,n(2)+r,\cdots\}}^{n}(\xi'_{i-d}P_{i-r-d}\xi_{i-d})^{1/2}(\xi'_{i-r-d}P_{i-r-d}\xi_{i-r-d})^{1/2}|\eta_i||y_{i-r}-\hat{y}_{i-r}|\\ &\leqq\sum_{i\leqq n}\xi'_{i-d}P_{i-r-d}\xi_{i-d}\eta_i^2/2+\sum_{i\leqq n}\xi'_{i-r-d}P_{i-r-d}\xi_{i-r-d}(y_{i-r}-\hat{y}_{i-r})^2/2.\end{aligned}\tag{5.21}$$

Applying (5.15), Lemma 1(ii), and Lemma 4(ii) to (5.20), and combining the result with (5.19), (5.20), (5.15), and (5.16), we obtain (5.18).

Suppose that $\sup_n|\varepsilon_n|<\infty$ almost surely. Then $\sup_n|\eta_n|<\infty$ almost surely and therefore

$$\sum_{j:n(j)\leqq n}\sum_{r=1}^{d-1}\|\xi_{n(j)+r-d}\||\eta_{n(j)+r}|=O\left(\sum_{j:n(j)\leqq n}\sum_{s=1}^{d-1}\|\xi_{n(j)-s}\|\right)\quad\text{a.s.}\tag{5.22}$$

Now assume that the $n(j)$ are stopping times with respect to $\{\mathscr{F}_{t-d+1}\}$. Then for $r=1,\cdots,d-1$, $T_j\triangleq n(j)+r-d$ is a stopping time with respect to $\{\mathscr{F}_t\}$ and $\eta_{n(j)+r}=\sum_{i=1}^{d}f_{d-i}\varepsilon_{T_j+i}(f_0=0)$. Moreover, $\|\xi_{n(j)+r-d}\|$ is $\mathscr{F}_{T_j}$-measurable. Hence an application of Lemma 1(iii) and (i) gives

$$\begin{aligned}&\sum_{j:n(j)\leqq n}\sum_{r=1}^{d-1}\|\xi_{n(j)+r-d}\||\eta_{n(j)+r}|\\ &=\sum_{j:n(j)\leqq n}\sum_{r=1}^{d-1}\|\xi_{n(j)+r-d}\|E\left(\left|\sum_{i=1}^{d}f_{d-i}\varepsilon_{T_j+i}\right|\,|\,\mathscr{F}_{T_j}\right)\\ &\quad+o\left(\sum_{j:n(j)\leqq n}\sum_{r=1}^{d-1}\|\xi_{n(j)+r-d}\|^2\right)+O(1)\quad\text{a.s.}\end{aligned}\tag{5.22'}$$

Since $E(|\sum_{i=1}^{d} f_{d-i}\varepsilon_{T_j+i}| \,|\, \mathcal{F}_{T_j}) \leqq E^{1/\alpha}(|\sum_{i=1}^{d} f_{d-i}\varepsilon_{T_j+i}|^{\alpha} \,|\, \mathcal{F}_{T_j})$, either (5.22′) or (5.22) implies that

$$\begin{aligned}&\sum_{j:n(j)\leqq n}\sum_{r=1}^{d-1}\|\xi_{n(j)+r-d}\|\,|\eta_{n(j)+r}|\\ &\qquad = O\Bigg(\sum_{j:n(j)\leqq n}\sum_{s=1}^{d-1}[\|\xi_{n(j)-s}\|+\|\xi_{n(j)-s}\|^2]\Bigg)+O(1) \quad \text{a.s.}\end{aligned} \tag{5.23}$$

Applying (5.15)-(5.18) and (5.23) to (5.10), we obtain that on E_1,

$$\begin{aligned}&Q_n+(1-2\lambda^2+o(1))\sum_{i=n(1)+1}^{n}[\xi_{i-d}'(\theta_{i-1}-\theta)]^2\\ &\quad\leqq(2d-1)\Big\{\limsup_{i\to\infty}E(\eta_i^2\,|\,\mathcal{F}_{i-d})+o(1)\Big\}\log\det P_{n-d}^{-1}\\ &\qquad + o\Bigg(\sum_{j:n(j)\leqq n}\sum_{r=1}^{h+d-1}[\|\xi_{n(j)-r}\|+\|\xi_{n(j)-r}\|^2]\Bigg)+O(1) \quad \text{a.s.}\end{aligned} \tag{5.24}$$

Since $Q_n\geqq 0$, it follows from (5.24) and (5.16) that on E_1

$$\begin{aligned}\sum_{i=n(1)}^{n}(\tilde{y}_i-\hat{y}_i)^2 &= \sum_{i=n(1)}^{n}(y_i-\hat{y}_i-\eta_i)^2\\ &\leqq(1+\lambda^2+o(1))(1-2\lambda^2+o(1))^{-1}(2d-1)\\ &\quad\times\Big\{\limsup_{i\to\infty}E(\eta_i^2\,|\,\mathcal{F}_{i-d})+o(1)\Big\}\log\det P_{n-d}^{-1}\\ &\quad + o\Bigg(\sum_{j:n(j)\leqq n}\sum_{r=1}^{h+d-1}[\|\xi_{n(j)-r}\|+\|\xi_{n(j)-r}\|^2]\Bigg)+o\Bigg(\sum_{i=1}^{n-d}\rho_i\Bigg)\\ &\quad +O(1) \quad \text{a.s.}\end{aligned} \tag{5.25}$$

Since λ can be arbitrarily small, (5.25) implies that (5.8) holds on $E_1\cap\{\log\det P_n^{-1}\to\infty\}=E\cap\{\xi_n'P_n\xi_n\to 0$ and $\lambda_{\max}(P_n^{-1})\to\infty\}$. Since $P(E)=1$ by (5.7), the desired conclusion follows. □

COROLLARY 4. *With the same notation and assumptions as in Theorem* 3(i), *suppose that the input-output data satisfy*

$$\sum_{i=1}^{n}(y_i^2+u_i^2)=O(n) \quad \textit{and} \quad \max_{i\leqq n}(y_i^2+u_i^2)=o(\log n) \quad \textit{a.s.} \tag{5.26}$$

Then

$$\sum_{i=1}^{n}\|\xi_i\|^2=O(n) \quad \textit{and} \quad \|\xi_n\|^2=o(\log n) \quad \textit{a.s.} \tag{5.27}$$

Suppose furthermore that $\rho_n=1/n$ *and that the stopping times* $n(1)<n(2)<\cdots$ *are so chosen that*

$$\sum_{j:n(j)\leqq n}\max_{1\leqq r\leqq h+d-1}(\|\xi_{n(j)-r}\|+\|\xi_{n(j)-r}\|^2)=O(\log n) \quad \textit{a.s.} \tag{5.28}$$

(*which is possible in view of* (5.27)). *Then* (5.1) *holds.*

Proof. Since $\tilde{y}_i=y_i-\eta_i$ and since $\sup_i|\varepsilon_i|<\infty$ almost surely, it follows from (5.26) that

$$\sum_{i=1}^{n}\|\psi_i\|^2=O(n) \quad \text{and} \quad \max_{i\leqq n}\|\psi_i\|^2=o(\log n) \quad \text{a.s.,} \tag{5.29}$$

where ψ_i is defined in (1.9). By Lemma 2,

$$\begin{aligned} C(q^{-1})(y_n-\hat{y}_n-\eta_n) &= -(\theta_{n-d}-\theta)'(\phi_{n-d}-\psi_{n-d}) \\ &\quad -(\theta_{n-d}-\theta)'\psi_{n-d}, \quad \text{and therefore} \\ \hat{C}_{n-d}(q^{-1})(y_n-\hat{y}_n-\eta_n) &= -(\theta_{n-d}-\theta)'\psi_{n-d}, \end{aligned} \tag{5.30}$$

noting that

$$\phi_{n-d}-\psi_{n-d} = -(0,\cdots,0,y_{n-1}-\hat{y}_{n-1}-\eta_{n-1},\cdots,y_{n-h}-\hat{y}_{n-h}-\eta_{n-h})'. \tag{5.31}$$

Since $\theta_{n-d}\to\theta$ almost surely, it follows from (5.29), (5.30), and Lemma 5(i) that

$$\sum_{i=1}^{n}(y_i-\hat{y}_i-\eta_i)^2 = O\left(\sum_{i=1}^{n}\|\theta_{i-d}-\theta\|^2\|\psi_{i-d}\|^2\right) = o(n) \quad \text{a.s.}, \tag{5.32}$$

$$y_n-\hat{y}_n-\eta_n = O\left(\max_{i\leqq n}\|\theta_{i-d}-\theta\|\,\|\psi_{i-d}\|\right) = o((\log n)^{1/2}) \quad \text{a.s.} \tag{5.33}$$

In view of (5.31), it follows from (5.29), (5.32), and (5.33) that

$$\sum_{i=1}^{n}\|\phi_i\|^2 = O(n) \quad \text{and} \quad \max_{i\leqq n}\|\phi_i\|^2 = o(\log n) \quad \text{a.s.} \tag{5.34}$$

Since $\hat{C}_n(q^{-1})\xi_n=\phi_n$ by (5.6b), (5.27) follows from (5.34) and Lemma 5(ii).

Suppose that $\rho_n=1/n$ and that (5.28) holds. Since $P_n^{-1}=P_{n(1)}^{-1}+\sum_{j=n(1)+1}^{n}\xi_j\xi_j'+\sum_{j=n(1)+1}^{n}I/j$, $\lambda_{\min}(P_n^{-1})\geqq(1+o(1))\log n(\to\infty)$ and $\xi_n'P_n\xi_n\leqq\|\xi_n\|^2/\lambda_{\min}(P_n^{-1})\to 0$ almost surely by (5.27). Since $\log\det P_n^{-1}\leqq(p(d)+k+d-1+h)\log\lambda_{\max}(P_n^{-1})$, the desired conclusion (5.1) follows from (5.8), (5.27), and (5.28). □

Theorems 3 and 2 have recently enabled us to provide an asymptotically efficient adaptive control rule in the general delay case and without stability assumptions on $A(z)$, (cf. [23]). The rule involves parallel implementation of the stochastic gradient and monitored recursive maximum likelihood algorithms. The stochastic gradient component of the controller serves to stabilize the system even when $A(z)$ is not stable, as an application of Theorem 2. Together with an occasional dither signal to perturb the target values, it also leads to well-excited blocks of input-output data from which strongly consistent recursive estimates of the system parameters can be obtained by the method of moments to guide the recursive maximum likelihood algorithm, giving a control rule that can be shown by an application of Theorem 3 to satisfy

$$\begin{aligned} &\limsup_{n\to\infty}\frac{\sum_{i=\Delta+1}^{n}(y_i-y_i^*-\eta_i)^2}{\log n} \\ &\qquad \leqq(2d-1)(p(d)+k+d-1+h)\limsup_{i\to\infty}E(\eta_i^2\,|\,\mathscr{F}_{i-d}) \quad \text{a.s.} \end{aligned} \tag{5.35}$$

in the general delay case and without assuming $A(z)$ to be stable (cf. [23]).

Instead of the preceding implicit approach, it is natural to ask whether similar results can be obtained by the explicit approach, involving the monitored recursive maximum likelihood estimator Θ_n of the parameter vector $\Theta=(-a_1,\cdots,-a_p, b_1,\cdots,b_k,c_1,\cdots,c_h)'$ for system (1.1):

$$\Theta_n=\pi_n(\Theta_{n-1}+P_{n-1}\zeta_{n-1}e_n), \tag{5.36a}$$

$$\begin{aligned} &\zeta_n+\hat{c}_{n,1}\zeta_{n-1}+\cdots+\hat{c}_{n,h}\zeta_{n-h}=\Phi_n \quad \text{where} \\ &\Phi_n=(y_n,\cdots,y_{n-p+1},u_{n-\Delta+1},\cdots,u_{n-\Delta-k+2},e_n,\cdots,e_{n-h+1})', \end{aligned} \tag{5.36b}$$

$$e_n=y_n-\Theta_{n-1}'\Phi_{n-1}, \tag{5.36c}$$

$$P_n^{-1}=P_{n-1}^{-1}+\zeta_n\zeta_n'. \tag{5.36d}$$

The π_n in (5.36a) is the projection, with respect to the norm induced by $P_{n-1}^{-1}=P_0^{-1}+\sum_{i=1}^{n-1}\zeta_i\zeta_i'$, into a closed convex subset $S_{n(j)}$ which can be conveniently chosen to be a cube in R^{p+k+h}. An analogue of Theorem 3 for the adaptive one-step ahead predictors $\hat{y}_{n+1}=\Theta_n'\Phi_n$ can be proved by similar (and simpler) arguments and is given in the following.

THEOREM 4. *Suppose that $C(z)$ is stable and that the random disturbances ε_n in the linear stochastic system* (1.1) *satisfy assumption* (1.3). *Let $n(1)<n(2)<\cdots$ be stopping times with respect to $\{\mathscr{F}_t\}$ and let $S_{n(j)}$ be an $\mathscr{F}_{n(j)}$-measurable, closed, and convex subset of R^{p+k+h} such that $P\{\Theta\in S_{n(j)}$ for all large $j\}=1$, where we define Θ and Ψ_t by* (1.11) *and* (1.12). *Define the monitored recursive maximum likelihood estimator Θ_n by* (5.36), *where π_n is given by* (5.5) (*with $d=1$ and $S_{n(j)}\subset R^{p+k+h}$*) *for $n(j)\leqq n<n(j+1)$. Then on $\{\lambda_{\max}(P_n^{-1})\to\infty$ and $\zeta_n'P_n\zeta_n\to 0\}$,*

$$\sum_{i\leqq n}(\Theta'\Psi_i-\Theta_i'\Phi_i)^2\leqq\left\{\limsup_{i\to\infty}E(\varepsilon_i^2|\mathscr{F}_{i-1})+o(1)\right\}\log\det P_n^{-1} + o\left(\sum_{j:n(j)\leqq n+1}\ \sum_{r=1}^{h}\|\zeta_{n(j)-r}\|^2\right)\quad a.s. \tag{5.37}$$

For $d>1$, analysis of the adaptive d-step ahead predictors defined from Θ_n using the explicit approach (1.18) becomes prohibitively difficult because of the inherent nonlinearities in (1.18) and (1.4), and it is doubtful that they would provide sharp results of the kind given by the implicit approach in Corollary 4.

REFERENCES

[1] K. J. ÅSTRÖM, *Introduction to Stochastic Control*, Academic Press, New York, 1970.

[2] L. LJUNG AND T. SÖDERSTRÖM, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, 1983.

[3] G. C. GOODWIN AND K. S. SIN, *Adaptive Filtering, Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[4] P. E. CAINES, *Linear Stochastic Systems*, John Wiley, New York, 1988.

[5] L. LJUNG, *On positive real transfer functions and the convergence of some recursive schemes*, IEEE Trans. Automat. Control, 22 (1977), pp. 539-551.

[6] ———, *Analysis of recursive stochastic algorithms*, IEEE Trans. Automat. Control, 22 (1977), pp. 551-575.

[7] J. B. MOORE AND G. LEDWICH, *Multivariable adaptive parameter and state estimators with convergence analysis*, J. Austral. Math. Soc. Ser. B, 21 (1979), pp. 176-197.

[8] V. SOLO, *The convergence of* AML, IEEE Trans. Automat. Control, 24 (1979), pp. 958-962.

[9] T. L. LAI AND C. Z. WEI, *Extended least squares and their applications to adaptive control and prediction in linear systems*, IEEE Trans. Automat. Control, 31 (1986), pp. 898-906.

[10] T. L. LAI, *Extended stochastic Lyapunov functions and recursive algorithms in linear stochastic systems*, in Stochastic Differential Systems: Proceedings of the 4th Bad Honnef Conference, N. Christopeit et al., eds., Springer-Verlag, New York, Berlin, Heidelberg, 1989, pp. 206-220.

[11] G. C. GOODWIN, P. J. RAMADGE, AND P. E. CAINES, *A globally convergent adaptive predictor*, Automatica—J. IFAC, 17 (1981), pp. 135-140.

[12] K. S. SIN, G. C. GOODWIN, AND R. B. BITMEAD, *An adaptive d-step ahead predictor based on least squares*, IEEE Trans. Automat. Control, 25 (1980), pp. 1161-1165.

[13] Y.-H. ZHANG, *Stochastic adaptive control and prediction based on modified least squares—the general delay-colored noise case*, IEEE Trans. Automat. Control, 27 (1982), pp. 1257-1260.

[14] G. C. GOODWIN, P. J. RAMADGE, AND P. E. CAINES, *Discrete time stochastic adaptive control*, SIAM J. Control Optim., 19 (1981), pp. 829-853.

[15] G. C. GOODWIN, K. S. SIN, AND K. K. SALUJA, *Stochastic adaptive control and prediction—the general delay-colored noise case*, IEEE Trans. Automat. Control, 25 (1980), pp. 946-949.

[16] J. J. FUCHS, *Indirect stochastic adaptive control: the general delay-white noise case*, IEEE Trans. Automat. Control, 27 (1982), pp. 219-223.

[17] ———, *Indirect stochastic adaptive control: the general delay-colored noise case*, IEEE Trans. Automat. Control, 27 (1982), pp. 470-472.

[18] T. L. Lai, C. Z. Wei, and Y. G. Zhang, *Convergence properties of some recursive identification schemes and adaptive predictors*, Proc. First Amer. Control Conference, Arlington, VA, 1982, pp. 176-180.

[19] T. L. Lai and C. Z. Wei, *Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems*, Ann. Statist., 10 (1982), pp. 154-166.

[20] H. J. Kushner and D. S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, New York, Berlin, Heidelberg, 1978.

[21] K. S. Sin and G. C. Goodwin, *Stochastic adaptive control using a modified least squares algorithm*, Automatica—J. IFAC, 18 (1982), pp. 315-321.

[22] M. S. Bazaraa and C. M. Shetty, *Nonlinar Programming—Theory and Algorithms*, John Wiley, New York, 1979.

[23] T. L. Lai and Z. Ying, *Parallel recursive algorithms in asymptotically efficient adaptive control of linear stochastic systems*, SIAM J. Control Optim., 29 (1991), pp. 1091-1126.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 5, pp. 1091-1127, September 1991

005

PARALLEL RECURSIVE ALGORITHMS IN ASYMPTOTICALLY EFFICIENT ADAPTIVE CONTROL OF LINEAR STOCHASTIC SYSTEMS*

TZE LEUNG LAI† AND ZHILIANG YING‡

Abstract. First, a review of some recent developments in stochastic adaptive control of linear stochastic systems is given. By integrating and refining several basic ideas in these developments, a relatively complete asymptotic solution to the adaptive control problem is then provided for such systems. The solution involves parallel implementation of a few basic recursive identification algorithms that serve to complement each other. It also involves occasional use of a dither signal to probe the system for information.

Key words. stochastic adaptive control, linear systems, self-optimizing, self-timing, asymptotic efficiency, stochastic approximation, recursive maximum likelihood, method of moments, occasional excitation, general delay, colored noise, martingale theory

AMS(MOS) subject classifications. primary 93E10, 93E20, 62L20; secondary 60G42

1. Introduction. A widely used stochastic model in the time series and stochastic control systems literature is the ARMAX system (autoregressive moving average system with exogenous inputs) defined by the linear difference equation

$$A(q^{-1})y_n = B(q^{-1})u_{n-d} + C(q^{-1})\varepsilon_n, \tag{1.1}$$

where $\{y_n\}$, $\{u_n\}$, and $\{\varepsilon_n\}$ denote the output, input, and disturbance sequences, respectively, $d \geqq 1$ represents the delay and

$$\begin{aligned} &A(q^{-1}) = 1 + a_1 q^{-1} + \cdots + a_p q^{-p}, \qquad B(q^{-1}) = b_1 + \cdots + b_k q^{-(k-1)}, \\ &C(q^{-1}) = 1 + c_1 q^{-1} + \cdots + c_h q^{-h} \end{aligned} \tag{1.2}$$

are scalar polynomials in the backward shift operator q^{-1}. Throughout the following we assume that $b_1 \neq 0$. Because of its theoretical interest and practical importance, the problem of determining the inputs u_n to keep the outputs y_{n+d} as close as possible to certain target values y^*_{n+d} when the system parameters are not known in advance but have to be estimated "on-line" (i.e., during the operation of the system) has been an important topic in the subject of stochastic adaptive control. Åström [1] and Kumar [2] have provided comprehensive surveys of the developments concerning this problem up to the early 1980s. In this paper, we first review several recent results and then integrate and refine some basic ideas in the literature to develop a relatively complete asymptotic solution to the adaptive control problem for ARMAX systems.

Throughout the following let $x_0 = (y_0, \cdots, y_{1-p}, u_0 \cdots u_{2-d-k}, \varepsilon_0, \cdots, \varepsilon_{1-h})'$ denote the "initial condition" of (1.1). Letting $\mathscr{G}_0$ be the σ-field generated by x_0 and letting $\mathscr{G}_n$ be the σ-field generated by $\{x_0, \varepsilon_1, \cdots, \varepsilon_n\}$, it will be assumed that

$$\begin{aligned} &\{\varepsilon_n, \mathscr{G}_n, n \geqq 1\} \text{ is a martingale difference sequence such that } E(\varepsilon_n^2 \mid \mathscr{G}_{n-1}) = \sigma^2 \\ &(\text{nonrandom}) > 0 \text{ and } \sup_n E(|\varepsilon_n|^\alpha \mid \mathscr{G}_{n-1}) < \infty \text{ almost surely for some } \alpha > 2. \end{aligned} \tag{1.3}$$

The input u_n is assumed to be "nonanticipating" in the sense that it involves only the current and past observations $y_n, y_{n-1}, u_{n-1}, \cdots$ and possibly also some extraneous

* Received by the editors November 21, 1989; accepted for publication (in revised form) July 26, 1990. This work was supported by the National Science Foundation, the National Security Agency, and the Air Force Office of Scientific Research.

† Department of Statistics, Stanford University, Stanford, California 94305.

‡ Department of Statistics, University of Illinois, Champaign, Illinois 61820.

randomization ω_n. Specifically, for $n \geqq 1$, we assume that u_n is measurable with respect to the σ-field $\mathscr{F}_n$ generated by $\{x_0, y_1, \cdots, y_n, u_1, \cdots, u_{n-1}, \omega_1, \cdots, \omega_n\}$, where

(1.4) ω_n is independent of $\{x_0, \varepsilon_1, \varepsilon_2, \cdots\} \cup \{\omega_1, \cdots, \omega_{n-1}\}$.

In view of (1.1), (1.3), and (1.4), an induction argument shows that

(1.5) $\mathscr{G}_n \subset \mathscr{F}_n$ and $\{\varepsilon_n, \mathscr{F}_n, n \geqq 1\}$ is a martingale difference such that $E(\varepsilon_n^2 | \mathscr{F}_{n-1}) = \sigma^2$ and $\sup_n E(|\varepsilon_n|^\alpha | \mathscr{F}_{n-1}) < \infty$ almost surely.

We call the polynomial $B(z) = b_1 + \cdots + b_k z^{k-1}$ *stable* if all its zeros lie outside the unit circle. We will also call two or more polynomials *coprime* (or *relatively prime*) if their greatest common divisors have degree 0.

In principle, given a (joint) probability distribution of the random sequence $\{x_0, \varepsilon_1, \varepsilon_2, \cdots\}$ and a prior distribution π of the unknown parameter vector

$$\theta = (-a_1, \cdots, -a_p, b_1, \cdots, b_k, c_1, \cdots, c_h)', \tag{1.6}$$

the problem of adaptive control of system (1.1) with quadratic costs is simply a dynamic programming problem in which the "state" is the conditional distribution of the original system state and parameter vector given the past observations (cf. [2]). In particular, we can use backward induction to solve the dynamic programming equations for inputs $u_1, \cdots, u_{N-d}$ that minimize

$$\int E_\theta\left\{\sum_{i=d+1}^{N} (y_i - y_i^*)^2\right\} d\pi(\theta), \tag{1.7}$$

for every given horizon N, where the y_i^* are nonrandom target values for the outputs. However, the dynamic programming equations are prohibitively difficult to handle, both computationally and analytically. Moreover, the need to specify a realistic probability law for the initial condition x_0 and the unobservable random errors ε_n together with a reasonable prior distribution for the parameter vector θ in this Bayesian approach may also cause difficulties in practice.

Despite these difficulties in the implementation of the Bayesian approach, Bayesian analysis of some very simple examples has provided important insights into the structure of optimal control rules. In particular, Feldbaum [3] and subsequent authors have shown that Bayes rules have the "dual control" function of both probing the system for information about its parameters and trying to drive the outputs toward their target values (cf. the review and references in § 5 of [2]). Åström [1, p. 478] has provided an interesting numerical example, which took 180 CPU hours to compute on a VAX 11/780 computer, to illustrate this dual control effect in the adaptive control problem (1.7) for the simple ARX model

$$y_n - \alpha y_{n-1} = \beta u_{n-1} + \varepsilon_n, \tag{1.8}$$

with independently and identically distributed zero-mean normal ε_n, a normal prior distribution for β, known value $\alpha = 1$, and $N = 30$, $y_i^* \equiv 0$. The example shows that the Bayes rule takes relatively large and irregular control actions to probe the system when the Bayes estimate $\hat{\beta}_t$ of β has poor precision, but is well approximated when $\hat{\beta}_t$ has high precision by the "certainty-equivalence" rule

$$\hat{\beta}_t u_t = -y_t. \tag{1.9}$$

Suppose that in Åström's example concerning the ARX model (1.8), instead of assuming α to be known and β to have a normal prior distribution, we assume $\beta \neq 0$ to be known and α to have a normal prior distribution. This problem turns out to be more tractable and can be used to derive lower bounds on the cost of Bayes rules. First, note that for any nonanticipating input sequence $\{u_n\}$,

$$\begin{aligned}\int_{-\infty}^{\infty} E_\alpha\left\{\sum_{i=2}^{N}(y_i-\varepsilon_i)^2\right\}d\pi(\alpha) &= \sum_{i=1}^{N-1} E\{(\alpha y_i+\beta u_i)^2\}\\ &\geqq \sum_{i=1}^{N-1} E\{[\alpha - E(\alpha\,|\,y_1,u_1,\cdots,y_i,u_i)]^2 y_i^2\}.\end{aligned} \tag{1.10}$$

In particular, for a nonanticipating input sequence $\{u_n\}$ that satisfies the property

$$\lim_{N\to\infty} N^{-1}\sum_{i=2}^{N}(y_i-\varepsilon_i)^2=0 \quad \text{a.s. } [P_\alpha] \quad \text{for every } \alpha, \tag{1.11}$$

it can be shown that

$$\begin{aligned}&\sum_{i=1}^{n}\{\alpha-E(\alpha\,|\,y_1,u_1,\cdots,y_i,u_i)\}^2y_i^2\\ &\quad=(1+o(1))\sum_{i=1}^{n}\left\{\sum_{t=2}^{i}y_{t-1}\varepsilon_t\Big/\sum_{t=2}^{i}y_{t-1}^2\right\}^2y_i^2+O(1)\sim\sigma^2\log n \text{ a.s. } [P_\alpha]\end{aligned} \tag{1.12}$$

for every α. The first relation in (1.12) makes use of the representation of the Bayes estimate $E(\alpha\,|\,y_1,u_1,\cdots,y_i,u_i)$ in the Gaussian model as a convex combination of the prior mean of α and the least squares estimate of α based on the data $\{y_1,u_1,\cdots,y_i,u_i\}$ (cf. [4, p. 32]). Noting that (1.11) implies that $\sum_1^N y_i^2\sim\sigma^2N$ and $\sup_{i\leqq N}y_i^2=o(N)$ almost surely $[P_\alpha]$, the second asymptotic relation in (1.12) can be proved by a partial summation argument similar to that of [5, pp. 1206-1207]. From (1.10) and (1.12), it follows by an application of Fatou's lemma and Fubini's theorem that

$$\int_{-\infty}^{\infty}E_\alpha\left\{\sum_{i=2}^{N}(y_i-\varepsilon_i)^2\right\}d\pi(\alpha)\geqq(1+o(1))\sigma^2\log N \tag{1.13}$$

for all nonanticipating input sequences $\{u_n\}$ satisfying (1.11).

Lai [4] has extended the results (1.10), (1.12), and (1.13) above for the regulation problem (1.7) (with $y_i^*\equiv 0$) to general ARX models (1.1) (with $C(q^{-1})=1$) and unit delay $d=1$. Specifically, assuming $b_1(\neq 0)$ to be known and putting a truncated normal prior distribution π on $\lambda=b_1^{-1}(a_1,\cdots,a_p,-b_2,\cdots,-b_k)'$, it is shown in [4] that in analogy with (1.13),

$$\int E_\lambda\left\{\sum_{i=2}^{N}(y_i-\varepsilon_i)^2\right\}d\pi(\lambda)\geqq(1+o(1))\sigma^2(p+k-1)\log N \tag{1.14}$$

for all nonanticipating input sequences $\{u_n\}$ satisfying (1.11) and the additional growth condition that $u_n^2=O(n^\delta)$ almost surely for some $0<\delta<1$. The truncated normal prior distribution π in (1.14) is the restriction of a standard multivariate normal distribution to the λ-region defined by the following:

(1.15) $A(z)$ and $B(z)$ are stable, and the polynomials $a_1z^{p-1}+\cdots+a_p$ and $z^{k-1}B(z^{-1})$ are relatively prime.

In the case of unit delay $d=1$, if all the parameters of system (1.1) are known, then the optimal controller chooses the input u_t at stage t so that $E(y_{t+1}\,|\,\mathcal{F}_t)=y_{t+1}^*$, and its output at stage $t+1$ is $y_{t+1}^* + \varepsilon_{t+1}$. In view of this, Lai [4] defines the "regret" at stage N of an input sequence $\{u_n\}$ to be

$$R_N = \sum_{i=2}^{N} \{y_i - (y_i^* + \varepsilon_i)\}^2 = \sum_{i=2}^{N} \{E(y_i \mid \mathcal{F}_{i-1}) - y_i^*\}^2. \tag{1.16}$$

Note that (1.14) above represents a lower bound for the expected regret in the regulation problem $y_i^* \equiv 0$ within a Bayesian framework. For general delay d, definition (1.16) of "regret" can be extended to

$$\begin{aligned} R_N &= \sum_{i=d+1}^{N} \{E(y_i \mid \mathcal{F}_{i-d}) - y_i^*\} = \sum_{i=d+1}^{N} (y_i - y_i^* - \eta_i)^2, \text{ where} \\ \eta_i &= y_i - E(y_i \mid \mathcal{F}_{i-d}) = \varepsilon_i + f_1\varepsilon_{i-1} + \cdots + f_{d-1}\varepsilon_{i-d+1} \text{ for } i \geqq d, \end{aligned} \tag{1.17}$$

with $f_1, \cdots f_{d-1}$ depending only on $A(z)$ and $C(z)$, as will be shown in § 2. The regret R_N, which is 0 for the optimal controller assuming knowledge of all system parameters so that u_t is determined by $E(y_{t+d} \mid \mathcal{F}_t) = y_{t+d}^*$, can be regarded as the cumulative cost up to stage N due to lack of knowledge of the system parameters in an adaptive controller. Since

$$E\left\{ \sum_{i=d+1}^{N} (y_i - y_i^*)^2 \right\} = E(R_N) + E\left(\sum_{i=d+1}^{N} \eta_i^2 \right), \tag{1.18}$$

the problem of minimizing the total expected quadratic cost is equivalent to that of minimizing the expected value of the regret R_N.

An input sequence $\{u_n\}$ is said to be "self-optimizing" (or "globally convergent") if

$$R_n / n \to 0 \quad \text{a.s.} \tag{1.19}$$

Using martingale theory (cf. Lemma 1 of § 2) and the identity $\sum_{d+1}^{n} (y_i - y_i^*)^2 = \sum_{d+1}^{n} \eta_i^2 + \sum_{d+1}^{n} (y_i - y_i^* - \eta_i)^2 + 2\sum_{d+1}^{n} (y_i - y_i^* - \eta_i)\eta_i$, it can be shown that (1.19) is equivalent to

$$n^{-1} \sum_{d+1}^{n} (y_i - y_i^*)^2 \to \sigma_d^2 \triangleq \sigma^2(1 + f_1^2 + \cdots + f_{d-1}^2) = E\eta_d^2 \quad \text{a.s.} \tag{1.19$'$}$$

Since σ_d^2 is the long-run average cost of the optimal controller defined by $E(y_{t+d} \mid \mathcal{F}_t) = y_{t+d}^*$ that assumes knowledge of the system parameters, the equivalence between (1.19) and (1.19′) explains the term "self-optimizing" (cf. [2], [6]).

The idea of working with the "regret" instead of the original cost criterion in opimization problems with unknown parameters was introduced by Lai and Robbins [5] in the context of choosing the design levels to minimize $E(\sum_1^N y_i^2)$ in the two-parameter regression model

$$y_n = \beta(u_{n-1} - \mu) + \varepsilon_n, \tag{1.20}$$

with independently and identically distributed normal errors ε_n such that $E\varepsilon_1 = 0$ and $E\varepsilon_1^2 = \sigma^2$. If μ is known, then $u_t \equiv \mu$ is the optical choice of the design levels. However, this experimental design does not provide any information about the unknown slope β. To resolve the dilemma between the control objective of setting the inputs near μ and the need for information to estimate β and μ, Lai and Robbins [5] started by considering the situation in which $\beta \neq 0$ is known. In this case, the maximum likelihood estimate based on $u_0, y_1, \cdots, u_{n-1}, y_n$ is $\hat{\mu}_n = n^{-1} \sum_{i=1}^{n} (u_{i-1} - \beta^{-1} y_i)$, and therefore $\hat{\mu}_n - \mu = -n^{-1} \sum_{i=1}^{n} \varepsilon_i / \beta$ irrespective of how the design levels u_i are chosen. This suggests that there is no conflict between information (to estimate μ) and control (setting u_i at the current best guess of μ). In particular, the certainty-equivalence rule $u_t=\hat{\mu}_t$ has regret

$$R_n=\sum_{i=2}^{n}(y_i-\varepsilon_i)^2=\sum_{i=1}^{n-1}\left(i^{-1}\sum_{t=1}^{i}\varepsilon_t\right)^2\sim\sigma^2\log n\quad\text{a.s.}\tag{1.21}$$

(cf. [5]). Moreover, if we put a normal prior distribution π on μ then the separation principle shows that the Bayes rule minimizing $\int_{-\infty}^{\infty}E_\mu(R_n)\,d\pi(\mu)$ is simply given at every stage t by

$$u_t=E(\mu\,|\,u_0,y_1,\cdots,u_{n-1},y_n)=\hat{\mu}_t+O(t^{-1}),\tag{1.22}$$

and therefore as in (1.21), the regret of this Bayes rule also satisfies

$$R_n\sim\sigma^2\log n\quad\text{a.s.}\tag{1.23}$$

(cf. [4]). Without assuming $\beta(\neq 0)$ to be known in advance and without assuming the ε_i to be normal, Lai and Robbins [5] made use of the theory of adaptive stochastic approximation to find an input sequence $\{u_n\}$ whose regret R_n also has the same logarithmic order (1.23) as the Bayes rule that assumes β to be known and the ε_i to be normal.

Note that (1.23) clearly implies the self-optimizing property (1.19). For the regression model (1.20), it is in fact very easy to construct self-optimizing inputs by using schemes with "forced" learning (cf. [2, p. 348]). We need only make sure that there is enough information to estimate β and μ consistently by setting the design levels at K_1 during stages $n_1<n_2<\cdots$ and at $K_2(\neq K_1)$ during stages $n_1^*<n_2^*<\cdots$, where $\{n_i\}$ and $\{n_i^*\}$ are disjoint sequences of positive integers such that $n_i/i\to\infty$ and $n_i^*/i\to\infty$. Let $\hat{\beta}_t$ and $\hat{\alpha}_t$ be the least squares estimates of the slope β and the intercept $-\beta\mu$ based on all the observations at times $n_i\leqq t$ and $n_i^*\leqq t$. The inputs $u_t=-(\hat{\alpha}_t/\hat{\beta}_t)I_{\{\hat{\beta}_t\neq 0\}}$ are self-optimizing, noting that the times n_i and n_i^* at which "probing inputs" are introduced occur infrequently. However, the fact that $\hat{\alpha}_t$ and $\hat{\beta}_t$ are based only on observations at these infrequent times suggests that the self-optimizing rule may be quite inefficient since we are wasting the information from the other observations. Therefore, not only should an asymptotically efficient control rule have the self-optimizing property (1.19), but it should also be able to attain the logarithmic order (1.23) for the regret, emulating the Bayes rules that assume knowledge of β.

Because of the dynamical structure defined by the linear difference equation (1.1), the problem of adaptive control of system (1.1) is much more complex than that of the regression model (1.20). For example, if $A(z)$ is not stable, then using white-noise probing inputs may lead to exponentially divergent output trajectories. In fact, the important and challenging problem of finding self-optimizing control schemes for (1.1) that can be implemented in real time has been an active area of research since the seminal paper of Åström and Wittenmark [7] on "self-tuning regulators."

The "self-tuning" idea is to start by considering the case where the system parameters are known, for which the optimal controller can be represented in some convenient recursive form, and then to replace the parameters in the optimal controller by recursive estimates that converge. Åström and Wittenmark [7] showed that if the recursive estimates, which they chose to be of least squares type, should converge to some limit, then substituting the parameter vector in the optimal controller by this limit must necessarily give the optimal controller, justifying the use of the adjective "self-tuning." A central problem with this approach, which still remains unsettled, is whether the least-squares-type estimates are indeed convergent. Goodwin, Ramadge, and Caines [8] circumvented this problem by introducing another method, called the "stochastic gradient" (stochastic approximation) algorithm, to estimate a linear transformation θ^* of the parameter vector θ, and were able to establish the self-optimizing property (1.19′) for their scheme. In § 2 we give a brief review of several ideas in these and subsequent developments.

Sections 2 and 5 also review some recent results on d-step ahead adaptive predictors constructed by the so-called direct (or implicit) method. These results enable us to unify, integrate, and extend two different approaches of constructing self-tuning control rules. The first approach, to be considered in § 3, uses a scalar gain in the recursive identification algorithm, as in the Goodwin–Ramadge–Caines [8] method and subsequent modifications thereof. We will focus on one such modification, proposed by Caines and Lafortune [9], that adds a dither signal to the control action to persistently excite the input-output process for consistent estimation of the system parameters. In § 3, we give a short proof of the main results of [8] and [9] by using a different argument, which also enables us to substantially generalize the algorithms of [8] and [9]. In particular, we can modify the continually disturbed control scheme of Caines and Lafortune [9] by only adding the dither signal occasionally, giving an "ocasionally excited" input-output process. Using the input-output data only from those stages at which the system is excited by the inclusion of the dither signal, we show in § 4 that strongly consistent estimators of the system parameters can be obtained by the method of moments.

As noted by Sin and Goodwin [10], although the global convergence proofs in [8] only apply to the stochastic approximation scheme, "it seems that in practically all applications of stochastic adaptive control, a least squares iteration is used" since "it generally has much superior rates of convergence compared with stochastic approximation." Instead of a scalar gain as in [8], a matrix gain such as in the Åström–Wittenmark scheme is typically used. In § 5 we consider this alternative approach that involves a matrix gain to construct self-tuning controllers. In particular, by using a consistent but inefficient estimator such as that developed in § 4 to monitor a recursive maximum likelihood estimator of the parameters in a reparametrized model, d-step ahead adaptive predictors associated with the recursive identification scheme are shown not to differ very much from the minimum variance predictors that assume knowlege of the system parameters. This result also suggests the possibility of achieving a logarithmic order for the regret (see (1.24) below) by using the certainty-equivalence rule associated with the monitored recursive maximum likelihood algorithm to define the inputs when certain conditions are met.

In § 6 we show that by a parallel implementation of stochastic approximation (involving a scalar gain) and monitored recursive maximum likelihood (involving a matrix gain), a self-tuning controller that integrates both methods can be constructed. The stochastic approximation component of the controller serves to stabilize the system even when $A(z)$ is not stable. Together with an occasional dither signal to perturb the target values, it also leads to excited blocks of input-output data from which strongly consistent estimators of the system parameters can be obtained by the method of moments to guide the recursive maximum likelihood algorithm. The monitored recursive maximum likelihood component of the controller can be shown to satisfy the conditions of the main theorem of § 5. Not only is the resultant self-tuning controller self-optimizing in the sense of (1.19), but it also attains the following logarithmic order for the regret:

$$(1.24) \qquad R_n \leqq (1+o(1))\sigma_d^2\{(p \vee (h-d+1))+h+k+d-1\}(2d-1)\log n \quad \text{a.s.},$$

where σ_d^2 is defined in (1.19′). Here and in the sequel we use $\vee$ and $\wedge$ to denote maximum and minimum, respectively. Moreover, if $\log(1+\sum_{i=1}^{n} y_i^{*2}) = o(\log n)$, as in the regulation problem ($y_i^* \equiv 0$), we can further strengthen (1.24) into

$$(1.25)\qquad R_n \leqq (1+o(1))\sigma_d^2\{(p \vee (h-d+1))+d+k-2\}(2d-1)\log n \quad \text{a.s.}$$

Note that in the case where $d=1$ and $C(q^{-1})=1$ (so that $h=0$), (1.25) reduces to

$$(1.26)\qquad R_n \leqq (1+o(1))\sigma^2(p+k-1)\log n \quad \text{a.s.}$$

Comparing (1.26) with the lower bound (1.14) for the expected value of R_n within a Bayesian framework that assumes knowledge of b_1 in advance, we see that the *certainty-equivalence* rule thus constructed *closely emulates* the performance of the Bayes rule. Lai [4] calls such control rules "asymptotically efficient," in analogy with the minimal order of magnitude (1.23) for the regret R_n in adaptive control of the simple regression model (1.20). Extensions of these to multivariable systems will be considered in § 7, which also contains several other concluding remarks.

2. Adaptive prediction, recursive identification, and some basic lemmas. To begin with, consider the unit-delay, white noise case (i.e., $d=1=C(q^{-1})$), for which (1.1) can be written as a stochastic regression model

$$(2.1)\qquad y_{n+1} = \theta'\psi_n + \varepsilon_{n+1},$$

where $\theta = (-a_1, \cdots, -a_p, b_1, \cdots, b_k)'$ and $\psi_n = (y_n, \cdots, y_{n-p+1}, u_n, \cdots, u_{n-k+1})'$. When θ is known, the minimum variance 1-step ahead predictor of y_{t+1} is $E(y_{t+1}|\mathcal{F}_t) = \theta'\psi_t$, and the optimal controller chooses u_t so that $\theta'\psi_t = y_{t+1}^*$. An obvious modification of this control rule for the case of unknown θ is to "adapt" the optimal predictor $\theta'\psi_t$ by substituting the unknown θ by the least squares estimate θ_t, which has the recursive representation

$$(2.2a)\qquad \theta_t = \theta_{t-1} + P_{t-1}\psi_{t-1}(y_t - \theta'_{t-1}\psi_{t-1}),$$

$$(2.2b)\qquad P_t = P_{t-1} - P_{t-1}\psi_t\psi'_t P_{t-1}/(1+\psi'_t P_{t-1}\psi_t).$$

A first difficulty with this straightforward approach is that the controller $\theta'_t\psi_t = y_{t+1}^*$ need not be well defined since the coefficient $b_1^{(t)}$ of u_t in $\theta'_t\psi_t$ may be 0 unless some continuity assumptions are made on the distribution of $\{\varepsilon_t\}$. For the regulation problem $y_i^* \equiv 0$, Åström and Wittenmark [7] circumvented this difficulty by reparametrizing (2.1) as

$$(2.3)\qquad \begin{aligned} &y_{n+1} = b_1(u_n - \lambda' X_n) + \varepsilon_{n+1}, \quad \text{where} \\ &\lambda = (a_1/b_1, \cdots, a_p/b_1, -b_2/b_1, \cdots, -b_k/b_1)', \\ &X_n = (y_n, \cdots, y_{n-p+1}, u_{n-1}, \cdots, u_{n-k+1})', \end{aligned}$$

and replacing the unknown b_1 by $b \neq 0$ while using the least squares criterion to estimate λ. They also showed that if the estimates should converge then they must converge to λ. However, a difficult open problem is whether with positive probability these estimates may fail to converge in their certainty-equivalence regulator

$$(2.4)\qquad b(u_t - \lambda'_t X_t) = 0, \quad \text{where } \lambda_t = \arg\min_\lambda \sum_{i=1}^{t} (y_i - bu_{i-1} + b\lambda' X_{i-1})^2.$$

Instead of adhering to a prior guess b of b_1, an obvious modification of the Åström-Wittenmark approach is to update this guess with the current and past data. Lai and Wei [11] recently considered this modification under the assumption that $A(z)$

is stable. They also introduced occasional blocks of white-noise probing inputs and used only the data from these stages of forced learning to obtain a strongly consistent estimate $b(t)$ of b_1. Replacing b by $b(t)$ in (2.4) whenever $b(t) \neq 0$ and the number of probing inputs up to stage t does not fall below some threshold K_t (with $K_t \to \infty$ and $K_t = o(\log t)$) leads to a certainty-equivalence regulator, for system (1.1) with $d = 1$ and $C(q^{-1}) = 1$, whose regret has the logarithmic order (1.26), as shown by Lai and Wei [11]. A key step in their proof is the analysis of the cumulative squared difference $\sum_1^n (\tilde{y}_{t+1} - \hat{y}_{t+1})^2$ between the optimal predictor $\tilde{y}_{t+1} = \theta'\psi_t = b_1(u_t - \lambda' X_t)$ and the adaptive predictor $\hat{y}_{t+1} = b(t)(u_t - \lambda_t' X_t)$ in the stochastic regression model (2.3).

For the unit-delay colored-noise case, (1.1) can still be written in the form of a stochastic regression model (2.1) with θ given by (1.6) and

$$\psi_n = (y_n, \cdots, y_{n-p+1}, u_n, \cdots u_{n-k+1}, \varepsilon_n, \cdots, \varepsilon_{n-h+1})'. \tag{2.5}$$

However, the regressor ψ_n contains unobservable components $\varepsilon_n, \cdots, \varepsilon_{n-h+1}$. Replacing the unobservable ε_i by their estimates $\hat{\varepsilon}_i$ in the recursion (2.2) leads to the "extended least squares" algorithm of the form

$$\theta_t = \theta_{t-1} + P_{t-1}\phi_{t-1}(y_t - \theta_{t-1}'\phi_{t-1}), \qquad P_t^{-1} = P_{t-1}^{-1} + \phi_t\phi_t', \tag{2.6}$$

$$\phi_t = (y_t, \cdots, y_{t-p+1}, u_t, \cdots, u_{t-k+1}, \hat{\varepsilon}_t, \cdots, \hat{\varepsilon}_{t-h+1})'. \tag{2.7}$$

The estimates $\hat{\varepsilon}_i$ of ε_i in (2.7) are given either by the residuals $\hat{\varepsilon}_i = y_i - \theta_i'\phi_{i-1}$, in which case (2.6) is called the AML algorithm, or by the prediction errors $\hat{\varepsilon}_i = y_i - \theta_{i-1}'\phi_{i-1}$, in which case (2.6) is called the RML1 algorithm. Assuming that

$$\operatorname{Re}(1/C(e^{it}) - 1/2) > 0 \text{ for all } t \in [-\pi, \pi], \tag{2.8}$$

Lai and Wei [12] showed that for the AML algorithm θ_n and its associated adaptive 1-step ahead predictor $\hat{y}_{n+1} = \theta_n'\phi_n$,

$$\begin{gathered} \|\theta_n - \theta\| = O(\{\log \lambda_{\max}(P_n^{-1})\}^{1/2}/\lambda_{\min}^{1/2}(P_n^{-1})) \quad \text{a.s.}, \\ \sum_1^n \|\phi_i - \psi_i\|^2 = O(\log \lambda_{\max}(P_n^{-1})) \quad \text{a.s.}, \end{gathered} \tag{2.9}$$

$$\sum_{i=1}^n (\theta'\psi_i - \theta_i'\phi_i)^2 I_{\{\phi_i' P_i \phi_i \leqq \delta\}} = O(\log \lambda_{\max}(P_n^{-1})) \quad \text{a.s. for every } 0 < \delta < 1. \tag{2.10}$$

Making use of (2.9) and (2.10), they proved not only the self-optimizing property (1.19) but also the much stronger conclusion $R_n = O(\log n)$ almost surely for a modification of the certainty-equivalence rule based on the AML algorithm, under the assumptions of boundedness of the target values y_i^*, stability of the open-loop plant (i.e., $A(z)$ and $B(z)$ are stable), and assumption (2.8) for $C(z)$. A basic ingredient of this modification is a simple criterion to decide whether information is inadequate for approximating the unobservable $\theta'\psi_t$ by $\theta_t'\phi_t$. When the data show inadequate information for such approximation, instead of adhering to the certainty-equivalence formula $\theta_t'\phi_t = y_{t+1}^*$ to determine the output u_t, [12] proposes to introduce a block of white noise perturbations to improve the information content of the design, in such a way that the number of these perturbations up to stage n is kept within the order $O(\log n)$. This approach ensures that $\theta_t'\phi_t$ is eventually close to $\theta'\psi_t$ whenever certainty-equivalence inputs are used, although θ_t may not converge to θ. In fact, "self-tuning" holds in the sense that

$$\sum_{t \leqq n: \theta_t'\phi_t = y_{t+1}^*} (\theta'\psi_t - \theta_t'\phi_t)^2 = O(\log n) \quad \text{a.s.} \tag{2.11}$$

For this case of general delay d, it is convenient to reparametrize (1.1) in the following prediction form. By the division algorithm, there exist polynomials $F(z)=1+f_1z+\cdots+f_{d-1}z^{d-1}$ and $G(z)=g_1+\cdots+g_{p(d)}z^{p(d)-1}$with $p(d)=p\vee(h-d+1)$ such that

$$C(z)=A(z)F(z)+z^dG(z), \tag{2.12}$$

and therefore (1.1) can be rewritten in the form

$$C(q^{-1})\{y_{n+d}-F(q^{-1})\varepsilon_{n+d}\}=G(q^{-1})y_n+B(q^{-1})F(q^{-1})u_n \tag{2.13}$$

(cf. [2, p. 368], [13, p. 268], [14, p. 134]). Hence, in the case of known system parameters, the optimal d-step ahead predictor $\tilde{y}_{n+d}\triangleq E(y_{n+d}\,|\,\mathcal{F}_n)$ is given by

$$\tilde{y}_{n+d}+c_1\tilde{y}_{n+d-1}+\cdots+c_h\tilde{y}_{n+d-h}=G(q^{-1})y_n+(BF)(q^{-1})u_n, \tag{2.14}$$

where $(BF)(z)=B(z)F(z)=b_1+(bf)_2z+\cdots+(bf)_{k+d-1}z^{k+d-2}$, and its prediction error is

$$\eta_{n+d}\triangleq y_{n+d}-\tilde{y}_{n+d}=F(q^{-1})\varepsilon_{n+d}. \tag{2.15}$$

Let

$$\tilde{\theta}=(g_1,\cdots,g_{p(d)},b_1,(bf)_2,\cdots,(bf)_{k+d-1},-c_1,\cdots,-c_h)'. \tag{2.16}$$

Since $y_{n+d}=\tilde{y}_{n+d}+\eta_{n+d}$, we have by (2.14) the following prediction form of (1.1):

$$\begin{aligned} &y_{n+d}=\tilde{\theta}'\tilde{\psi}_n+\eta_{n+d}, \quad \text{where}\\ &\tilde{\psi}_n=(y_n,\cdots,y_{n-p(d)+1},u_n,\cdots,u_{n-k-d+2},\tilde{y}_{n+d-1},\cdots,\tilde{y}_{n+d-h})'. \end{aligned} \tag{2.17}$$

In the case of known system parameters, the optimal controller chooses u_n so that $\tilde{y}_{n+d}=y^*_{n+d}$. Without assuming knowledge of the parameters, it is natural to replace $\tilde{y}_{n+d}=\tilde{\theta}'\tilde{\psi}_n$ in the optimal controller by an adaptive predictor $\hat{y}_{n+d}$. In particular, Goodwin, Ramadge, and Caines [8] defined $\hat{y}_{n+d}$ by replacing the $\tilde{y}_i$ in $\tilde{\psi}_n$ by the target values y^*_i and the unknown parameters $\tilde{\theta}$ in (2.16) by a stochastic gradient (stochastic approximation) estimate of the form

$$\theta^*_n=\theta^*_{n-d}+(a/r^*_{n-d})\phi^*_{n-d}(y_n-\theta^{*\prime}_{n-d}\phi^*_{n-d}), \tag{2.18a}$$

$$r^*_n=r^*_{n-d}+\|\phi^*_n\|^2, \tag{2.18b}$$

where $\phi^*_n=(y_n,\cdots,y_{n-p(d)+1},u_n,\cdots,u_{n-k-d+2},y^*_{n+d-1},\cdots,y^*_{n+d-h})'$ and u_n is defined by

$$\theta^{*\prime}_n\phi^*_n=y^*_{n+d}. \tag{2.18c}$$

To ensure that the coefficient $b^*_{1,n}$ of u_n in (2.18c) is nonzero almost surely, Goodwin, Ramadge, and Caines assume that

(2.19) $(x_0,\varepsilon_1,\cdots,\varepsilon_n)$ is absolutely continuous with respect to Lebesgue measure for every $n\geqq 1$.

Their use of the scalar gain $1/r^*_{n-d}$ in (2.18) instead of the matrix gain $(P_0^{-1}+\sum_1^{n-d}\phi^*_i\phi^{*\prime}_i)^{-1}$ as in (2.2) or (2.6) considerably simplifies the analysis. Under certain additional stability and positive real assumptions on $B(z)$ and $C(z)$ and boundedness assumptions on $\{y^*_i\}$, they proved the self-optimizing property (1.19′) for the certainty-equivalence rule (2.18c) in the case $d=1$ and also in the case $C(q^{-1})=1$ and general d. For general delay and colored noise, Goodwin, Sin, and Saluja [15] suggested replacing (2.18b) by

$$r^*_n=r^*_{n-1}+\|\phi^*_n\|^2, \tag{2.20}$$

and showed how to extend the argument in [8] to prove the self-optimizing property of their scheme.

Clearly, other recursive identification algorithms can also be used to define adaptive predictors $\hat{y}_{n+d}$ and thereby to construct certainty-equivalence rules. The *direct* (or *implicit*) approach of adaptive prediction is to first obtain a recursive estimator θ_n of the parameter $\tilde{\theta}$ of the implicit system (2.17) and then to substitute the optimal predictor $\tilde{\theta}'\tilde{\psi}_n$ of y_{n+d} by $\theta'_n\phi_n$, where ϕ_n is a pseudoregression vector in which the undetermined components $\tilde{y}_i$ of $\tilde{\psi}_n$ are replaced by $\theta'_{i-d}\phi_{i-d}$, cf. [13, p. 181]. We have recently developed in [16] analogues of the bound (2.10) for the cumulative squared difference $\sum_1^n (\tilde{\theta}'\tilde{\psi}_i - \theta'_i\phi_i)^2$ between the optimal d-step ahead predictor $\tilde{\theta}'\tilde{\psi}_i$ and the adaptive predictor $\theta'_i\phi_i$ associated with a variety of recursive identification algorithms, including stochastic gradient, extended least squares, and monitored recursive maximum likelihood. In particular, for adaptive predictors associated with the stochastic gradient algorithm, the results of [16] are summarized in the following.

THEOREM 1. *Suppose that* $\operatorname{Re}\{C(e^{it})-(d-\frac{1}{2})a\}>0$ *for all* $|t|\leqq\pi$ *and some* $a>0$, *and that the random disturbances* ε_n *in the linear stochastic system* (1.1) *satisfy assumption* (1.3). *Let* $p(d)=p\vee(h-d+1)$. *Consider the stochastic gradient algorithm that estimates the parameter* $\tilde{\theta}$ *in the reparametrized model* (2.17) *by the recursion*

$$\theta_{n,G}=\theta_{n-1,G}+(a/r_{n-d})(y_n-\hat{y}_{n,G})\phi_{n-d,G}, \tag{2.21a}$$

$$\phi_{n,G}=(y_n,\cdots,y_{n-p(d)+1},u_n,\cdots,u_{n-k-d+2},\hat{y}_{n+d-1,G},\cdots,\hat{y}_{n+d-h,G})', \tag{2.21b}$$

$$r_n=r_{n-1}+\|\phi_{n,G}\|^2, \tag{2.21c}$$

$$\hat{y}_{n+d,G}=\theta'_{n,G}\phi_{n,G}. \tag{2.21d}$$

Then

$$\limsup_{n\to\infty}\|\theta_{n,G}\|<\infty \quad a.s., \tag{2.22}$$

$$\sum_{i=1}^{n}(\tilde{\theta}'\tilde{\psi}_i-\theta'_{i,G}\phi_{i,G})^2=o(r_n)+O(1) \quad a.s. \tag{2.23}$$

Moreover, if $B(z)$ *is stable, then*

$$\begin{gathered}\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}(\tilde{\theta}'\tilde{\psi}_i-\theta'_{i,G}\phi_{i,G})^2=0 \quad \textit{and}\\ \limsup_{n\to\infty} n^{-1}\sum_{i=1}^{n}(y_i^2+u_i^2)<\infty \quad \textit{a.s. on} \left\{\limsup_{n\to\infty} n^{-1}\sum_{i=1}^{n}(\theta'_{i,G}\phi_{i,G})^2<\infty\right\}.\end{gathered} \tag{2.24}$$

The proof of Theorem 1 and of analogous results for other recursive identification algorithms in [16] makes use of the idea of "extended stochastic Lyapunov functions." For an algorithm with a matrix gain P_{n-d} of the form $P_t^{-1}=P_{t-1}^{-1}+\phi_t\phi'_t$, [16] makes use of the extended stochastic Lyapunov function $(\theta_n-\tilde{\theta})'P_{n-d}^{-1}(\theta_n-\tilde{\theta})$, which typically does not converge. For the stochastic gradient algorithm with the scalar gain $1/r_{n-d}$, [16] uses $Q_n=\|\theta_{n,G}-\tilde{\theta}\|^2$ as the extended stochastic Lyapunov function and derives from the recursion (2.21a) recursive inequalities for Q_n. In the case of unit delay ($d=1$), Goodwin, Ramadge, and Caines [8] also derived similar recursive inequalities. However, in order to apply the martingale convergence theorem, they introduced a transformation of the form $Z_n=Q_n+S_n/r_{n-1}$ with $S_n\geqq 0$, and used the recursive inequalities to conclude that Z_n is a nonnegative almost supermartingale (or stochastic Lyapunov function; cf. [17], [18]). Instead of relying on the martingale convergence theorem, [16] works directly with the recursive inequalities for Q_n and applies certain martingale limit theorems, restated in Lemma 1 below for subsequent reference, to analyze these recursive inequalities, which turn out to involve not only Q_n but also $\sum_1^{n-d} (\tilde{\theta}'\tilde{\psi}_i - \theta_{i,G}\phi_{i,G})^2/r_i$. Thus Theorem 1 follows from such analysis of Q_n.

LEMMA 1. *Let $\{\varepsilon_n\}$ be a martingale difference sequence with respect to an increasing sequence of σ-fields $\{\mathscr{F}_n\}$ such that* $\sup_n E(|\varepsilon_n|^\alpha || \mathscr{F}_{n-1}) < \infty$ *almost surely for some $\alpha > 2$. Let z_n be an $\mathscr{F}_{n-1}$-measurable random variable for every n.*

(i) *$\sum_1^n z_i\varepsilon_i$ converges almost surely on $\{\sum_1^\infty z_i^2 < \infty\}$, and for every $\eta > \frac{1}{2}$,*

$$\left(\sum_1^n z_i\varepsilon_i\right) \Big/ \left(\sum_1^n z_i^2\right)^\eta \to 0 \quad a.s.\ on \left\{\sum_1^\infty z_i^2 = \infty\right\}.$$

Consequently,

$$\sum_1^n z_i\varepsilon_i = o\left(\sum_1^n z_i^2\right) + O(1) \quad a.s. \tag{2.25}$$

(ii) *$\sum_1^n |z_i|\varepsilon_i^2 = O(\sum_1^n |z_i|)$ almost surely on $\{\sup_n |z_n| < \infty\}$. Moreover,*

$$\sum_1^n |z_i|\varepsilon_i^2 = \sum_1^n |z_i|E(\varepsilon_i^2 | \mathscr{F}_{i-1}) + o\left(\sum_1^n |z_i|\right) \quad on \left\{\sup_n |z_n| < \infty, \sum_1^\infty |z_n| = \infty\right\}. \tag{2.26}$$

In §§ 3 and 6 we apply Theorem 1 and some extensions thereof to develop, for general delay and colored noise and without stability assumptions on $A(z)$, asymptotically efficient adaptive control schemes that satisfy (1.24) or (1.25) by using parallel implementation of $\theta_{n,G}$ and three other recursive identification algorithms. The remainder of this section states for subsequent reference some algebraic lemmas on the dynamics of (1.1) and on recursive estimation of its parameters.

LEMMA 2. *Suppose that for $n \leqq t < m$, $\phi_t = (y_t, \cdots, y_{t-\nu+1}, u_t, \cdots, u_{n-\kappa+1}, \hat{y}_{t+d-1}, \cdots, \hat{y}_{t+d-h})'$ and $C(q^{-1})(y_{t+d} - \eta_{t+d}) = G(q^{-1})y_t + \Gamma(q^{-1})u_t$, where $C(q^{-1}) = 1 + c_1q^{-1} + \cdots + c_hq^{-h}$, $G(q^{-1}) = g_1 + \cdots + g_\nu q^{-(\nu-1)}$ and $\Gamma(q^{-1}) = \gamma_1 + \cdots + \gamma_\kappa q^{-(\kappa-1)}$ are polynomials in the backward shift operator q^{-1}. Suppose that there exist $(\nu + \kappa + h) \times 1$ vectors θ_t such that $\hat{y}_{t+d} = \theta_t'\phi_t$ for $n \leqq t < m$. Then*

$$C(q^{-1})(y_s - \hat{y}_s - \eta_s) = -(\theta_{s-d} - \tilde{\theta})'\phi_{s-d} \quad for\ m + d > s \geqq n + d, \tag{2.27}$$

where $\tilde{\theta} = (g_1, \cdots, g_\nu, \gamma_1, \cdots, \gamma_\kappa, -c_1, \cdots, -c_h)'$.

LEMMA 3. *Let $\{D_n, n \geqq 0\}$ be a sequence of $L \times L$ real matrices such that $\sum_0^\infty \|D_n\| < \infty$ and let $D(z) = \sum_{n=0}^\infty D_nz^n$. Suppose that $D(e^{it}) + D'(e^{-it})$ is nonnegative definite for all $t \in [-\pi, \pi]$.*

(i) *Let $\{g_n, n \geqq 0\}$ be a sequence of $L \times 1$ real vectors and let $f_n = \sum_{k=0}^n D_kg_{n-k}$. Then for any $N \geqq 0$, $\sum_{n=0}^N f_n'g_n \geqq 0$.*

(ii) *Suppose that $D_i = 0$ for all $i > h$. Let $M \geqq h$ and suppose that $f_n = \sum_{j=0}^h D_jg_{n-j}$ for $M \leqq n \leqq N$. Let $\{r_n, M \leqq n \leq N\}$ be a nondecreasing sequence of positive numbers. Then*

$$\sum_{n=M}^N f_n'g_n/r_n \geqq \sum_{j=1}^h \sum_{t=0}^{h-j} g_{M-1-t}'D_{j+t}'g_{M-1+j}/r_{M-1+j}.$$

LEMMA 4. (i) *Consider the linear system $A(q^{-1})y_n = q^{-d}B(q^{-1})u_n + C(q^{-1})\varepsilon_n$, in which $A(q^{-1})$, $B(q^{-1})$, and $C(q^{-1})$ are the polynomials (1.2) in the unit delay operator q^{-1}. Suppose that $b_1 \neq 0$ and that $B(z)$ is stable. Then there exist $K > 0$, $0 < \rho < 1$, and $\alpha_0, \alpha_1, \cdots, \beta_0, \beta_1, \cdots$ such that* $\max(|\alpha_i|, |\beta_i|) \leqq K\rho^i$ *and for all $t \geqq n$,*

$$\left|u_t - \sum_{i=0}^{t-n} \alpha_iy_{t+d-i} - \sum_{i=0}^{t-n} \beta_i\varepsilon_{t+d-i}\right| \leqq K\rho^{t-n}\left\{\sum_{\nu=1}^{k-1} |u_{n-\nu}| + \sum_{\nu=1}^{p} |y_{n+d-\nu}| + \sum_{\nu=1}^{h} |\varepsilon_{n+d-\nu}|\right\}.$$

Consequently, there exists $K^*>0$ such that for all $t \geqq n+d$,

$$\sum_{i=n+d}^{t} u_{i-d}^2 \leqq K^* \left\{ \sum_{i=n-p+d}^{t} y_i^2 + \sum_{i=n-h+d}^{t} \varepsilon_i^2 + \sum_{\nu=1}^{k-1} u_{n-\nu}^2 \right\}.$$

(ii) *Suppose that the polynomial $C(z)=1+c_1z+\cdots+c_hz^h$ is stable. For $j=1,\cdots,h$, let $\{c_{n,j}\}$ be a sequence of numbers such that $\lim_{n\to\infty} c_{n,j}=c_j$, and let $C_n(q^{-1})=1+c_{n,1}q^{-1}+\cdots+c_{n,h}q^{-h}$. Suppose that ξ_n and ϕ_n are $L\times 1$ vectors such that $C_n(q^{-1})\xi_n=\phi_n$. Then there exist $K>0$ and $0<\rho<1$ such that for all $t>n$*

$$\|\xi_t\| \leqq K\left\{ \sum_{i=0}^{t-n-1} \rho^i \|\phi_{t-i}\| + \rho^{t-n} \sum_{r=0}^{h-1} \|\xi_{n-r}\| \right\}.$$

For the proofs of Lemmas 2 and 3, see [16]. The main ideas of the proof of Lemma 4(i) are given in [11, pp. 470-471], and the proof of Lemma 4(ii) uses similar arguments (cf. Step 1 in the Appendix of [12]).

LEMMA 5. *Let A_n be a symmetric, positive-definite $L\times L$ matrix. Suppose that*

$$\log \operatorname{tr}(A_n) \leqq (1+o(1)) \log n \text{ as } n\to\infty, \tag{2.28}$$

and that there exist $\tilde{L}(\leqq L)$ linearly independent $L\times 1$ vectors $v_1,\cdots,v_{\tilde{L}}$ satisfying the condition

$$\lim_{n\to\infty} \log(v_i'A_nv_i)/\log n = 0 \text{ for } i=1,\cdots,\tilde{L}. \tag{2.29}$$

Then $\limsup_{n\to\infty} (\log \det A_n)/\log n \leqq L-\tilde{L}$.

Proof. Let $\lambda_{n,1},\cdots,\lambda_{n,L}$ be the eigenvalues of the symmetric matrix A_n, and let $x_{n,1},\cdots,x_{n,L}$ be corresponding orthonormal eigenvectors. Let π_n be a permutation of $\{1,\cdots,L\}$ such that

$$\max_{1\leqq i\leqq \tilde{L}} |v_i'x_{n,\pi_n(1)}| \geqq \cdots \geqq \max_{1\leqq i\leqq \tilde{L}} |v_i'x_{n,\pi_n(L)}|. \tag{2.30}$$

We first show that

$$\rho \triangleq \liminf_{n\to\infty} \max_{1\leqq i\leqq \tilde{L}} |v_i'x_{n,\pi_n(\tilde{L})}| > 0. \tag{2.31}$$

Suppose that (2.31) is not true. Then in view of (2.30), there exist orthonormal vectors $X_{\tilde{L}},\cdots,X_L$ which are limit points of the sequences $\{x_{n,\pi_n(\tilde{L})}\},\cdots,\{x_{n,\pi_n(L)}\}$ such that $\max_{1\leqq i\leqq \tilde{L}} |v_i'X_j|=0$ for $\tilde{L}\leqq j\leqq L$. Hence the orthonormal vectors $X_{\tilde{L}},\cdots,X_L$ belong to the orthogonal complement of the linearly independent set $\{v_1,\cdots,v_{\tilde{L}}\}$, which is impossible since $(L-\tilde{L}+1)+\tilde{L}>L$.

Since $v_i'A_nv_i=\sum_{j=1}^{L}(v_i'x_{n,\tau_n(j)})^2\lambda_{n,\pi_n(j)} \geqq \sum_{j=1}^{\tilde{L}}(v_i'x_{n,\pi_n(j)})^2\lambda_{n,\pi_n(j)}$, it follows from (2.30) and (2.31) that for all large n,

$$\lambda_{n,\pi_n(j)} < 2\rho^{-2} \max_{1\leqq i\leq \tilde{L}} |v_i'A_nv_i|, \qquad j=1,\cdots,\tilde{L}. \tag{2.32}$$

From (2.32), it follows that for all large n,

$$\begin{aligned}\log\det A_n &= \sum_{j=1}^{L} \log \lambda_{n,\pi_n(j)} \leqq 2\rho^{-2}\tilde{L} \max_{1\leqq i\leqq \tilde{L}} \log(v_i'A_nv_i) + (L-\tilde{L}) \log \operatorname{tr} A_n \\ &\leqq (1+o(1))(L-\tilde{L}) \log n,\end{aligned}$$

by (2.28) and (2.29). □

LEMMA 6. *Consider the matrix polynomials (over the complex field)*

$$\begin{aligned} &A(z)=I+A_1z+\cdots+A_pz^p, \qquad B(z)=B_1+\cdots+B_kz^{k-1}, \\ &C(z)=I+C_1z+\cdots+C_hz^h, \end{aligned} \tag{2.33}$$

where I is the identity matrix and the A_i, B_i, and C_i are $\nu\times\nu$ matrices such that $\det B(z)\neq 0$ *for all* $|z|\leqq 1$. *Let* $\cdots, \varepsilon_{-1}, w_{-1}, \varepsilon_0, w_0, \varepsilon_1, w_1, \cdots$ *be independent, bounded* $\nu\times 1$ *random vectors such that the* ε_i *have a common distribution with mean* 0 *and positive-definite covariance matrix* $\sum$, *and the* w_i *have a common distribution with mean* 0 *and positive-definite covariance matrix V. Let* $d\geqq 1$ *and let* $\eta_i=\varepsilon_i+F_1\varepsilon_{i-1}+\cdots+F_{d-1}\varepsilon_{i-d+1}$, *where* $F_1,\cdots,F_{d-1}$ *are nonrandom* $\nu\times\nu$ *matrices. Let* $y_i=\eta_i+w_{i-d}$ *and define* u_i *by the linear difference equation*

$$A(q^{-1})y_n=B(q^{-1})u_{n-d}+C(q^{-1})\varepsilon_n, \tag{2.34}$$

where q^{-1} *denotes the backward shift operator. Suppose that the matrix polynomials* $z^pA(z^{-1})$, $z^{k-1}B(z^{-1})$, *and* $z^hC(z^{-1})$ *are left coprime, i.e., the determinants of their greatest common left divisors have degree* 0. *Then the* $\nu(p+k)\times\nu\{p+2(k-1)\nu+1\}$ *matrix H has full rank* $\nu(p+k)$, *where*

$$\begin{aligned}H=E\{&(y'_{-1},\cdots,y'_{-p},u'_{-d},\cdots,u'_{-d-k+1})'\\ &\times(y'_{-h-1},\cdots,y'_{-h-(k-1)\nu},w'_{-d},\cdots,w'_{-d-p-(k-1)\nu})\}.\end{aligned} \tag{2.35}$$

Proof. Let $b(z)=\det B(z)$ and let $\tilde{B}(z)$ denote the adjoint of $B(z)$. Then

$$\tilde{B}(z)B(z)=b(z)I,\quad \deg b(z)\leqq \nu(k-1),\quad \deg\tilde{B}(z)\leqq(\nu-1)(k-1). \tag{2.36}$$

Since $b(z)\neq 0$ for all $|z|\leqq 1$, we can invert (2.34) as

$$u_{n-d}=\{\tilde{B}(q^{-1})/b(q^{-1})\}\{A(q^{-1})y_n-C(q^{-1})\varepsilon_n\}. \tag{2.37}$$

Suppose that H is not of full rank. Then there exist $\nu\times 1$ vectors $\pi_1,\cdots,\pi_p$, $\gamma_0,\cdots\gamma_{k-1}$, not all 0, such that $(\pi'_1,\cdots,\pi'_p,\gamma'_0,\cdots,\gamma'_{k-1})H=0$, i.e.,

$$E\left\{\left(\sum_{i=1}^{p}\pi'_iy_{-i}+\sum_{i=0}^{k-1}\gamma'_iu_{-d-i}\right)y'_{-j}\right\}=0,\quad j=h+1,\cdots,h+(k-1)\nu, \tag{2.38}$$

$$E\left\{\left(\sum_{i=1}^{p}\pi'_iy_{-i}+\sum_{i=0}^{k-1}\gamma'_iu_{-d-i}\right)w'_{-j}\right\}=0,\quad j=d,\cdots,d+p+(k-1)\nu. \tag{2.39}$$

Let

$$L(z)=\sum_{i=1}^{p}\pi'_iz^i,\qquad \Gamma(z)=\sum_{i=0}^{k-1}\gamma'_iz^i. \tag{2.40}$$

Since $y_n=\eta_n+w_{n-d}$, it follows from (2.37) that

$$\begin{aligned}\sum_{i=1}^{p}\pi'_iy_{n-i}+\sum_{i=0}^{k-1}\gamma'_iu_{n-d-i}&=\{L(q^{-1})+\Gamma(q^{-1})\tilde{B}(q^{-1})A(q^{-1})/b(q^{-1})\}(\eta_n+w_{n-d})\\ &\quad-\{\Gamma(q^{-1})\tilde{B}(q^{-1})C(q^{-1})/b(q^{-1})\}\varepsilon_n.\end{aligned} \tag{2.41}$$

Recalling that $\{w_t\}$ and $\{\varepsilon_t\}$ are independent zero-mean random vectors with positive covariance matrices, (2.41) (with $n=0$) and (2.39) imply that

$$L(q^{-1})+\Gamma(q^{-1})\tilde{B}(q^{-1})A(q^{-1})/b(q^{-1})=q^{-(p+(k-1)\nu+1)}R(q^{-1})$$

for some power series $R(z)=\sum_{i=0}^{\infty}\rho_iz^i$, where the ρ_i are $1\times\nu$ vectors such that $\sum_0^\infty\|\rho_i\|<\infty$. Therefore

$$b(z)L(z)+\Gamma(z)\tilde{B}(z)A(z)=b(z)z^{p+(k-1)\nu+1}R(z). \tag{2.42}$$

In view of (2.36) and (2.40), the left-hand side of (2.42) is a matrix polynomial of degree less than or equal to $(k-1)\nu+p$, and therefore (2.42) implies that $R(z)=0$. Hence

$$b(q^{-1})L(q^{-1})=-\Gamma(q^{-1})\tilde{B}(q^{-1})A(q^{-1}). \tag{2.43}$$

From (2.41) and (2.43), it follows that

$$\sum_{i=1}^{p}\pi_i' y_{n-i}+\sum_{i=0}^{k-1}\gamma_i' u_{n-d-i}=-\{\Gamma(q^{-1})\tilde{B}(q^{-1})C(q^{-1})/b(q^{-1})\}\varepsilon_n=-\sum_{i=0}^{\infty}d_i'\varepsilon_{n-i}, \tag{2.44}$$

for some nonrandom $\nu\times 1$ vectors d_i such that $\sum_0^\infty \|d_i\|<\infty$. Recalling that $\operatorname{cov}(\varepsilon_n)$ is positive-definite and that $y_s=\varepsilon_s+F_1\varepsilon_{s-1}+\cdots+F_{d-1}\varepsilon_{s-d+1}+w_{s-d}$, it follows from (2.38) and (2.44) (with $n=0$) that

$$\Gamma(z)\tilde{B}(z)C(z)=b(z)\left\{\sum_{i=0}^{h}d_i'z^i+\sum_{i=h+(k-1)\nu+1}^{\infty}d_i'z^i\right\}. \tag{2.45}$$

In view of (2.36) and (2.40), the left-hand side of (2.45) is a polynomial of degree $\leqq k-1+(\nu-1)(k-1)+h<h+(k-1)\nu+1$, and therefore (2.45) implies that

$$\Gamma(z)\tilde{B}(z)C(z)=b(z)D(z)\quad\text{where } D(z)=\sum_{i=0}^{h}d_i'z^i. \tag{2.46}$$

From (2.40) and (2.43), $0=L(0)=-\Gamma(0)B_1^{-1}A(0)=-\gamma_0'B_1^{-1}$, noting that $\tilde{B}(0)/b(0)=(B(0))^{-1}=B_1^{-1}$ and that B_1 is invertible. Hence $\gamma_0=0$. Therefore $\Gamma(z)=\sum_{i=1}^{k-1}\gamma_i'z^i=z\sum_{j=0}^{k-2}\gamma_{j+1}'z^j$, $L(z)=z\sum_{j=0}^{p-1}\pi_{j+1}'z^j$. Setting $z=0$ in (2.46) then gives $(\det B(0))d_0'=0$, implying that $d_0=0$ and therefore $D(z)=z\sum_{j=0}^{h-1}d_{j+1}'z^j$. Hence (2.43) and (2.46) can be written as

$$\begin{gathered}\Gamma^*(z)\tilde{B}(z)A(z)/b(z)=-L^*(z),\ \Gamma^*(z)\tilde{B}(z)C(z)/b(z)=D^*(z),\text{ where}\\ L^*(z)=\sum_0^{p-1}\pi_{j+1}'z^j,\quad \Gamma^*(z)=\sum_0^{k-2}\gamma_{j+1}'z^j,\quad D^*(z)=\sum_0^{h-1}d_{j+1}'z^j.\end{gathered} \tag{2.47}$$

Let $\mathscr{B}(z)=z^{k-1}B(z^{-1})=B_1z^{k-1}+\cdots+B_k$, $\beta(z)=\det(\mathscr{B}(z))$ and let $\tilde{\mathscr{B}}(z)$ denote the adjoint of $\mathscr{B}(z)$. The $\beta(z)=z^{(k-1)\nu}\det(B(z^{-1}))$ and likewise $\tilde{\mathscr{B}}(z)=z^{(k-1)(\nu-1)}\tilde{B}(z^{-1})$, so

$$\tilde{B}(z^{-1})/b(z^{-1})=z^{k-1}\mathscr{B}(z)/\beta(z). \tag{2.48}$$

By (2.47) and (2.48),

$$\begin{gathered}\{z^{k-2}\Gamma^*(z^{-1})\}\{\tilde{\mathscr{B}}(z)/\beta(z)\}\{z^pA(z^{-1})\}=-z^{p-1}L^*(z^{-1}),\\ \{z^{k-2}\Gamma^*(z^{-1})\}\{\tilde{\mathscr{B}}(z)/\beta(z)\}\{z^hC(z^{-1})\}=z^{h-1}D^*(z^{-1}).\end{gathered}$$

Hence $\sum_1^{k-1}\gamma_{k-i}'z^{i-1}\{\tilde{\mathscr{B}}(z)/\beta(z)\}(z^pA(z^{-1}), z^hC(z^{-1}))$ is a matrix polynomial. This implies that $\gamma_1=\cdots=\gamma_{k-1}=0$ by Lemma 1(ii) of [19], since $\mathscr{B}(z)$, $z^pA(z^{-1})$, and $z^hC(z^{-1})$ are left coprime. Hence $\Gamma(z)=z\Gamma^*(z)=0$, and by (2.43), $L(z)=0$, contradicting that $\pi_1,\cdots,\pi_p,\gamma_0,\cdots,\gamma_{k-1}$ are not all zero. □

3. Stochastic gradient algorithms and occasionally disturbed control schemes. As an application of Theorem 1, we give a short proof of the self-optimizing property

(1.19) of the Goodwin–Ramadge–Caines scheme (2.18) in the unit-delay case ($d = 1, p(d) = p \vee h$), under their assumptions (2.19) and

$$B(z) \text{ is stable,} \tag{3.1a}$$

$$\operatorname{Re}\{C(e^{it}) - a/2\} > 0 \text{ for all } |t| \leqq \pi, \tag{3.1b}$$

$$\{y_t^*\} \text{ is a bounded nonrandom sequence.} \tag{3.1c}$$

In view of (2.18c),the vector ϕ_t^* in (2.18) is equivalent to (2.21b). Moreover, by (2.18c) and (3.1c), $\sup_n n^{-1} \sum_1^n (\theta_i^{*\prime}\phi_i^*)^2 < \infty$ almost surely. By (2.17) and (2.18c), $y_{n+1} - \varepsilon_{n+1} - y_{n+1}^* = \tilde{\theta}'\tilde{\psi}_n - \theta_n^{*\prime}\phi_n^*$, and therefore by Theorem 1, $n^{-1}\sum_1^n (y_i - y_i^* - \varepsilon_i)^2 \to 0$ almost surely, i.e., (1.19) holds for the Goodwin–Ramadge–Caines scheme (2.18).

To obtain strongly consistent estimates of the parameter vector (1.6) in the unit-delay case, Caines and Lafortune [9] replaced the definition of ϕ_i^* and (2.18c) in the Goodwin–Ramadge–Caines scheme by

$$\begin{aligned} &\theta_n^{*\prime}\phi_n^* = y_{n+1}^* + w_n, \\ &\phi_n^* = (y_n, \cdots, y_{n-p(1)+1}, u_n, \cdots u_{n-k+1}, y_n^* + w_{n-1}, \cdots, y_{n+1-h}^* + w_{n-h})', \end{aligned} \tag{3.2}$$

where w_t is independent of $\{x_0, \varepsilon_1, \varepsilon_2, \cdots, u_1, \cdots, u_{t-1}, w_1, \cdots, w_{t-1}\}$ and such that $Ew_t = 0$, $Ew_t^2 = v > 0$ and $\sup_t Ew_t^4 < \infty$. They proved that for this modified scheme

$$n^{-1}\sum_2^n (y_i - y_i^* - w_{i-1} - \varepsilon_i)^2 \to 0 \quad \text{a.s.} \tag{3.3}$$

Replacing y_i^* in the preceding paragraph by $y_i^* + w_{i-1}$, we can also obtain (3.3) as another corollary of Theorem 1. Making use of (3.3), Caines and Lafortune [9] showed that $n^{-1}\sum_{i=1}^n \psi_i\psi_i'$ converges almost surely to a positive-definite matrix. Using this "persistent excitation" property of the ψ_i and Solo's [20] consistency theorem for the AML algorithm, they obtained strongly consistent AML estimates of the system parameters. The control scheme (3.2), however, is not self-optimizing since (3.3) implies that $n^{-1}\sum_2^n (y_i - y_i^* - \varepsilon_i)^2 \to v > 0$ almost surely (cf. [9]). To preserve the self-optimizing property, the white-noise perturbations w_n should only be introduced in (3.2) infrequently, with relative frequency diminishing to 0, as will be done in Theorem 4 below.

The interlacing algorithm (2.18) is a composite of d recursions for $\theta_{j+nd}^*(n \geqq 1)$, $j = 0, 1, \cdots, d-1$. In contrast, the stochastic gradient algorithm of Theorem 1 does not involve such multiple recursions, and handles general delay d in exactly the same way as it handles the case $d = 1$. Thus, Theorem 1 leads to a stochastic gradient certainty-equivalence rule that differs from that of Goodwin et al. [8], [15] in the case where $d \geqq 2$ since interlacing is no longer needed.

The proof of Theorem 1 given in [16] can be readily extended to the situation in which the stochastic gradient scheme is applied only to broken blocks of successive observations and in which white-noise perturbations w_i are introduced only occasionally, an improvement over the Caines–Lafortune [9] continually disturbed control scheme. Moreover, our modification also works for general delay d without interlacing, and by choosing the perturbations w_i to have a continuous distribution, we can further dispense with the restrictive assumption (2.19) in the Caines–Lafortune scheme. This is the content of Theorem 2.

THEOREM 2. *Suppose that* $\operatorname{Re}\{C(e^{it}) - (d - \frac{1}{2})a\} > 0$ *for all* $|t| \leqq \pi$ *and some* $a > 0$, *and that the random disturbances* ε_n *in the linear stochastic system* (1.1) *satisfy assumption*

(1.3) with $\sup_n |\varepsilon_n| < \infty$ *almost surely. Let* $p(d) = p \vee (h - d + 1)$. *Let* $m_0 = 0$ *and let* $n_1 < m_1 < n_2 < m_2 < \cdots$ *be stopping times (with respect to* $\{\mathscr{F}_t\}$*) such that*

$$m_j - n_j \geqq d + h. \tag{3.4}$$

Let $\{w_n\}$ *be a sequence of random variables such that*

(3.5) w_n *is independent of* $\{x_0, \varepsilon_1, \varepsilon_2, \cdots, u_1, \cdots, u_{n-1}, w_1, \cdots, w_{n-1}\}$ *and the* w_n *are identically distributed with a common continuous distribution such that* $Ew_1 = 0$, $Ew_1^2 = v > 0$, *and* $|w_1| \leqq c$.

Define the modified stochastic gradient algorithm $\theta_{n,\tilde{G}}$ *recursively as follows: Choose* $\theta_{0,\tilde{G}}$ *such that its component* $b_{0,\tilde{G}}$ *estimating the component* b_1 *of the vector* $\tilde{\theta}$ *in (2.16) is nonzero. Let* $\theta_{t,\tilde{G}} = \theta_{0,\tilde{G}}$ *for* $t < n_1 + d$. *Let*

$$\phi_{n,\tilde{G}} = (y_n, \cdots, y_{n-p(d)+1}, u_n, \cdots, u_{n-k-d+2}, w_{n-1}, \cdots, w_{n-h})'. \tag{3.6}$$

For $m_{j-1} + d - 1 < n < n_j + d$, *define* $\theta_{n,\tilde{G}} = \theta_{m_{j-1}+d-1,\tilde{G}}$. *For* $n_j + d \leq n \leq m_j + d - 1$ *define*

$$\begin{aligned} \theta_{n,\tilde{G}} &= \theta_{n-1,\tilde{G}} + (a/\tilde{r}_{n-d})(y_n - w_{n-d})\phi_{n-d,\tilde{G}}, \\ \tilde{r}_{n-d} &= \tilde{r}_{n-d-1} + \|\phi_{n-d,\tilde{G}}\|^2, \end{aligned} \tag{3.7a}$$

$$\begin{aligned} \tilde{r}_{n_j-1} = {} & \sum_{i=1}^{j-1} \sum_{t=n_i+d}^{m_i+d-1} \|\phi_{t-d,\tilde{G}}\|^2 \\ & + \sum_{i=1}^{j} \left(1 \vee \sum_{\nu=0}^{p(d)-1} y_{n_i-\nu}^2 \vee \sum_{\nu=1}^{k+d-2} u_{n_i-\nu}^2\right) \\ & \times \log^2 \left\{ \sum_{i=1}^{j} \left(2 \vee \sum_{\nu=0}^{p(d)-1} y_{n_i-\nu}^2 \vee \sum_{\nu=1}^{k+d-2} u_{n_i-\nu}^2\right)\right\}. \end{aligned} \tag{3.7b}$$

Suppose that for $n_j \leqq n < m_j$ *the input* u_n *is chosen so that*

$$\theta'_{n,\tilde{G}}\phi_{n,\tilde{G}} = w_n \text{ if } b_{n,\tilde{G}} \neq 0, \quad \text{and } u_n = w_n \text{ otherwise,} \tag{3.8}$$

where $b_{n,\tilde{G}}$ *is the component of* $\theta_{n,\tilde{G}}$ *estimating* b_1. *Define* η_t *as in (2.15). Then* $P\{b_{n,\tilde{G}} = 0\} = 0$ *for every n and*

$$\limsup_{j\to\infty} \|u_{m_j+d-1,\tilde{G}}\| < \infty, \quad \lim_{j\to\infty} \sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} (y_t - \eta_t - w_{t-d})^2/\tilde{r}_{m_j-1} = 0 \quad a.s. \tag{3.9}$$

Proof. We first prove by induction that $P\{b_{n,\tilde{G}} = 0\} = 0$ for every n. By our choice of $\theta_{0,\tilde{G}}$, this is clearly true for $0(=m_0) \leqq n < n_1 + d$. Suppose that this is true for all $n \leqq t - 1$. If $n_j + d \leqq t \leqq m_j + d - 1$ for some j, then by (3.7a) and (3.8) together with (2.17),

$$\begin{aligned} b_{t,\tilde{G}} &= b_{t-1,\tilde{G}} + (a/\tilde{r}_{n-d})(\tilde{\theta}'\tilde{\psi}_{t-d} + \eta_t - w_{t-d})u_{t-d} \\ &= b_{t-1,\tilde{G}} + a(\tilde{r}_{t-d}b_{t-d,\tilde{G}})^{-1}(\tilde{\theta}'\tilde{\psi}_{t-d} + \eta_t - w_{t-d}) \\ &\quad \times \left\{ w_{t-d} - \sum_{i=0}^{p(d)-1} \lambda_i y_{t-d-i} - \sum_{i=1}^{k+d-2} \tilde{\lambda}_i u_{t-d-i} - \sum_{i=1}^{h} \lambda_i^* w_{t-d-i} \right\} \quad \text{a.s.,} \end{aligned} \tag{3.10}$$

where the λ_i, $\tilde{\lambda}_i$, λ_i^* represent the components of $\theta_{t-d,\tilde{G}}$. Since w_{t-d} is independent of $S_t \triangleq (\theta'_{0,\tilde{G}}, \cdots, \theta'_{t-1,\tilde{G}}, \tilde{\psi}'_{t-d}, x'_0, \tilde{r}_{t-d}, \eta_t, y_1, \cdots, y_{t-d}, u_1, \cdots, u_{t-d-1}, w_1, \cdots, w_{t-d-1})$ by (3.5) and (3.7), and since w_{t-d} has a continuous distribution, it then follows from (3.10) that $P\{b_{t,\tilde{G}} = 0 | S_t\} = 0$. If $m_{j-1} + d - 1 < t < n_j + d$ for some j, then $b_{t,\tilde{G}} = b_{m_{j-1},\tilde{G}}$, and by the induction assumption, $P\{b_{m_{j-1},\tilde{G}} = 0\} = 0$.

For simplicity, we will use θ_n and ϕ_n to denote $\theta_{n,\tilde G}$ and $\phi_{n,\tilde G}$. For $n_j \leqq n < m_j$, since $b_{n,\tilde G} \neq 0$ almost surely, (3.8) implies that $w_n = \theta'_n \phi_n$ almost surely. Therefore by Lemma 2, with probability 1,

$$C(q^{-1})(y_t - w_{t-d} - \eta_t) = -(\theta_{t-d} - \tilde\theta)'\phi_{t-d} \quad \text{for } n_j + d \leqq t \leqq m_j + d - 1. \tag{3.11}$$

Let $e_t = y_t - w_{t-d}$. From (3.7a) and (2.17), it follows by simple algebra (cf. the proof of Theorem 1 given in [16]) that

$$\begin{aligned} \|\theta_{n_{i+1}+d-1} - \tilde\theta\|^2 &= \|\theta_{m_i+d-1} - \tilde\theta\|^2 \\ &= \|\theta_{n_i+d-1} - \tilde\theta\|^2 + 2a \sum_{t=n_i+d}^{m_i+d-1} \{\tilde r_{t-d}^{-1}(e_t - \eta_t)\phi'_{t-d}(\theta_{t-1} - \tilde\theta) \\ &\quad + (a/2)\tilde r_{t-d}^{-2}(e_t - \eta_t)^2 \|\phi_{t-d}\|^2\} \\ &\quad + a^2 \sum_{t=n_i+d}^{m_i+d-1} \tilde r_{t-d}^{-2} \|\phi_{t-d}\|^2 \eta_t^2 + 2a \sum_{t=n_i+d}^{m_i+d-1} \eta_t \\ &\quad \cdot \{\tilde r_{t-d}^{-1}\phi'_{t-d}(\theta_{t-1} - \tilde\theta) + a\tilde r_{t-d}^{-2}(e_t - \eta_t)\|\phi_{t-d}\|^2\}. \end{aligned} \tag{3.12}$$

First note that $\{n_i + d \leqq t \leqq m_i + d - 1\} = \{n_i \leqq t - d\} \cap \{m_i > t - d\} \in \mathscr{F}_{t-d}$. Furthermore, ϕ_{t-d}, $\tilde r_{t-d}$ and $e_t - \eta_t = \tilde\theta' \tilde\psi_{t-d} - \theta'_{t-d}\phi_{t-d}$ are $\mathscr{F}_{t-d}$-measurable. Therefore an application of Lemma 1 (cf. the proof of Theorem 1 given in [16]) shows that with probability 1,

$$\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde r_{t-d}^{-2} \|\phi_{t-d}\|^2 \eta_t^2 = O\left(\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde r_{t-d}^{-2} \|\phi_{t-d}\|^2 \right) = O(1), \tag{3.13}$$

$$\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde r_{t-d}^{-2} \|\phi_{t-d}\|^2 (e_t - \eta_t)\eta_t = o\left(\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde r_{t-d}^{-2} (e_t - \eta_t)^2 \right) + O(1). \tag{3.14}$$

Writing $\theta_{t-1} - \tilde\theta = \sum_{s=1}^{d-1} (\theta_{t-s} - \theta_{t-s-1}) + (\theta_{t-d} - \tilde\theta)$ and noting that $\theta_{t-s} = \theta_{t-s-1}$ if $n_i \leqq t - s < n_i + d$, the same argument as that used in [16] for the proof of Theorem 1 can be used to show that with probability 1,

$$\begin{aligned} &\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde r_{t-d}^{-1} \phi'_{t-d}(\theta_{t-1} - \tilde\theta)\eta_t \\ &\quad = o\left(\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde r_{t-d}^{-2} [\phi'_{t-d}(\theta_{t-d} - \tilde\theta)]^2 \right) + o\left(\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde r_{t-d}^{-1} (e_t + \eta_t)^2 \right) \\ &+ O(1), \end{aligned} \tag{3.15}$$

$$\begin{aligned} &\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \{\tilde r_{t-d}^{-1}(e_t - \eta_t)\phi'_{t-d}(\theta_{t-1} - \tilde\theta) + (a/2)\tilde r_{t-d}^{-2}(e_t - \eta_t)^2 \|\phi_{t-d}\|^2\} \\ &\quad \leqq -(\rho + o(1)) \sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde r_{t-d}^{-1}(e_t - \eta_t)^2 + O(1) \\ &\qquad - \sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde r_{t-d}^{-1}(e_t - \eta_t)[\{C(q^{-1}) - (d - 1/2)a - \rho\}(e_t - \eta_t)], \end{aligned} \tag{3.16}$$

where $\rho > 0$ is so chosen that $\min_{|t| \leqq \pi} \operatorname{Re}\{C(e^{it}) - (d - 1/2)a - \rho\} > 0$.

From (3.11), it follows that

$$\begin{aligned} &\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde r_{t-d}^{-2} [\phi'_{t-d}(\theta_{t-d} - \tilde\theta)]^2 \\ &\quad = O\left(\sum_{i=1}^{j} \left\{ \sum_{t=n_i+d}^{m_i+d-1} \tilde r_{t-d}^{-2} (e_t - \eta_t)^2 + \tilde r_{n_i}^{-2} \max_{1 \leqq \nu \leqq h} (e_{n_i+d-\nu} - \eta_{n_i+d-\nu})^2 \right\} \right). \end{aligned} \tag{3.17}$$

Using the recursion $y_s = -\sum_{\nu=1}^{p} a_\nu y_{s-\nu} + \sum_{\nu=d}^{k+d-1} b_{\nu-d+1} u_{s-\nu} + \varepsilon_s + \sum_{\nu=1}^{h} c_\nu \varepsilon_{s-\nu}$, we can proceed inductively from $t=1$ to $t=(d-1)\vee 1$ to show that

$$\max_{1\leqq t\leqq d-1} y_{n_i+t}^2 = O\left(\sum_{\nu=0}^{p-1} y_{n_i-\nu}^2 \vee \sum_{\nu=1}^{k+d-2} u_{n_i-\nu}^2 \vee \sum_{\nu=1-d}^{h-1} \varepsilon_{n_i-\nu}^2\right),$$

in which the left-hand side is interpreted as 0 if $d-1=0$. Since $\sup_n |\varepsilon_n| < \infty$ almost surely, this implies that with probability 1,

$$\max_{1\wedge(d-h)\leqq t\leqq d-1} (y_{n_i+t}^2 + \eta_{n_i+t}^2) = O\left(1 \vee \nabla_{\nu=1}^{k+d-2} u_{n_i-\nu}^2 \vee \sum_{\nu=0}^{p(d)-1} y_{n_i-\nu}^2\right). \tag{3.18}$$

Recalling that $\sup_n |w_n| \leqq c$ and that $e_n = y_n - w_{n-d}$, we then obtain from (3.17) and (3.18) together with the definition of $\tilde{r}_{n_i-1}$ in (3.7b) that with probability 1,

$$\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde{r}_{t-d}^{-2} [\phi'_{t-d}(\theta_{t-d} - \tilde{\theta})]^2 = O\left(\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde{r}_{t-d}^{-1} (e_t - \eta_t)^2\right) + O(1). \tag{3.19}$$

Since $\operatorname{Re}\{C(e^{it}) - (d-1/2)a - \rho\} > 0$ for all $|t| \leqq \pi$, it follows from Lemma 3(ii) that

$$\begin{aligned} &-\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde{r}_{t-d}^{-1} (e_t - \eta_t) [\{C(q^{-1}) - (d-1/2)a - \rho\}(e_t - \eta_t)] \\ &\leqq -\sum_{i=1}^{j} \sum_{l=1}^{h} \sum_{\nu=0}^{h-1} c_{l+\nu} (e_{n_i+d-1-\nu} - \eta_{n_i+d-1-\nu})(e_{n_i+d-1+l} - \eta_{n_i+d-1+l}) / \tilde{r}_{n_i+l-1} \\ &\leqq \frac{1}{2} \sum_{i=1}^{j} \sum_{l=0}^{h-1} \sum_{\nu=1}^{h-1} \left\{ \frac{c_{l+\nu}^{1/2}}{\rho^{1/2} \tilde{r}_{n_i+l}^{1/2}} (e_{n_i+d-\nu} - \eta_{n_i+d-\nu}) \right\}^2 + \frac{\rho}{2} \sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i} \frac{(e_t - \eta_t)^2}{\tilde{r}_{t-d}}, \end{aligned} \tag{3.20}$$

using the inequality $|AB| \leqq (A^2 + B^2)/2$ and noting that $n_i + d + h \leqq m_i$. Moreover, since $e_n = y_n - w_{n-d}$ and $\sup_n (|w_n| + |\eta_n|) < \infty$ almost surely, it follows from (3.18) and (3.7b) that

$$\sum_{i=1}^{\infty} \sum_{\nu=1}^{h} (e_{n_i+d-\nu} - \eta_{n_i+d-\nu})^2 / \tilde{r}_{n_i} < \infty \quad \text{a.s.} \tag{3.21}$$

Combining (3.16) with (3.20) and (3.21) gives

$$\begin{aligned} &\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \{\tilde{r}_{t-d}^{-1} (e_t - \eta_t) \phi'_{t-d} (\theta_{t-1} - \tilde{\theta}) + (a/2) \tilde{r}_{t-d}^{-2} (e_t - \eta_t)^2 \|\phi_{t-d}\|^2\} \\ &\leqq -(\rho/2 + o(1)) \sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde{r}_{t-d}^{-1} (e_t - \eta_t)^2 + O(1) \quad \text{a.s.} \end{aligned} \tag{3.22}$$

From (3.12)-(3.15), (3.19), and (3.22), it follows that

$$\|\theta_{n_j+d-1} - \tilde{\theta}\|^2 \leqq -(\rho/2 + o(1)) \sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i+d-1} \tilde{r}_{t-d}^{-1} (e_t - \eta_t)^2 + O(1) \quad \text{a.s.}$$

Hence $\limsup_{j\to\infty} \|\theta_{m_{j-1}+d-1} - \tilde{\theta}\|^2 < \infty$ and $\sum_{i=1}^{\infty} \sum_{t=n_i+d}^{m_i+d-1} \tilde{r}_{t-d}^{-1} (e_t - \eta_t)^2 < \infty$ almost surely, implying the desired conclusion (3.9) by the Kronecker lemma. □

COROLLARY 1. *With the same notation and assumptions as in Theorem* 2, *let* $\#_j = \sum_{i=1}^{j} (m_i - n_i)$. *Assume furthermore that* $B(z)$ *is stable and that*

$$\begin{aligned} &\sum_{i=1}^{j} \left(1 \vee \sum_{\nu=0}^{p(d)-1} y_{n_i-\nu}^2 \vee \sum_{\nu=1}^{k+d-2} u_{n_i-\nu}^2\right) \log^2 \left\{\sum_{i=1}^{j} \left(2 \vee \sum_{\nu=0}^{p(d)-1} y_{n_i-\nu}^2 \vee \sum_{\nu=1}^{k+d-2} u_{n_i-\nu}^2\right)\right\} \\ &\qquad = O(\#_j) \quad \textit{a.s.} \end{aligned} \tag{3.23}$$

Then

$$\sum_{i=1}^{j}\left(\sum_{t=n_i}^{m_i+d-1} y_t^2+\sum_{t=n_i+d}^{m_i+d-1} u_{t-d}^2\right)=O(\#_j) \quad a.s., \tag{3.24}$$

$$\lim_{j\to\infty}\sum_{i=1}^{j}\sum_{t=n_i+d}^{m_i+d-1}(y_t-\eta_t-w_{t-d})^2/\#_j=0 \quad a.s. \tag{3.25}$$

Proof. From (3.6), (3.7b), and (3.23), it follows that

$$\tilde{r}_{m_j-1}=O\left(\#_j+\sum_{i=1}^{j}\sum_{t=n_i+d}^{m_i+d-1}(y_{t-d}^2+u_{t-d}^2)\right) \quad \text{a.s.} \tag{3.26}$$

Since $b_1\neq 0$ and $B(z)$ is stable, an application of Lemma 4(i) shows that

$$\sum_{t=n_i+d}^{m_i+d-1} u_{t-d}^2\leqq K\left\{\sum_{t=n_i-p+d}^{m_i+d-1} y_t^2+\sum_{t=n_i-h+d}^{m_i+d-1}\varepsilon_t^2+\sum_{\nu=1}^{k-1}u_{n_i-\nu}^2\right\} \tag{3.27}$$

for some $K>0$. In view of the assumption that $\sup_t|\varepsilon_t|<\infty$ almost surely, it follows from (3.26), (3.27), and (3.23) that

$$\tilde{r}_{m_j-1}=O\left(\#_j+\sum_{i=1}^{j}\sum_{t=n_i+1}^{m_i+d-1} y_t^2\right) \quad \text{a.s.} \tag{3.28}$$

Letting $e_t=y_t-w_{t-d}$, it follows from (3.9) and (3.28) that

$$\begin{aligned}\sum_{i=1}^{j}\sum_{t=n_i+d}^{m_i+d-1} e_t^2&\leqq 2\sum_{i=1}^{j}\sum_{t=n_i+d}^{m_i+d-1}\{\eta_t^2+(e_t-\eta_t)^2\}\\ &=O(\#_j)+o\left(\#_j+\sum_{i=1}^{j}\sum_{t=n_i+1}^{m_i+d-1} y_t^2\right) \quad \text{a.s.}\end{aligned} \tag{3.29}$$

Since $y_t^2\leqq 2e_t^2+2w_{t-d}^2$, the desired conclusion (3.24) follows from (3.29), (3.23), and (3.27). Moreover, from (3.9), (3.26), and (3.24), (3.25) follows. □

The proof of Theorem 1 given in [16] can also be readily extended to show that the same conclusions hold for the following modification of the stochastic gradient algorithm (2.21), noting that condition (3.32) below implies that ρ_n is $\mathcal{F}_{n-d}$-measurable and that $(\alpha+\rho_n w_n)^2\leqq \alpha a$, $(\rho_n\theta'_{n-d,G}\phi_{n-d,G})^2=O(r_{n-d})$. This modification can be used to define the input u_t by

$$\theta'_{t,G}\phi_{t,G}=y^*_{t+d}, \tag{3.30}$$

since $b_{t,G}\neq 0$ almost surely even without the continuity assumption (2.19) (cf. Theorem 2 and [21]).

THEOREM 3. *Suppose that in Theorem* 1 *we change* (2.21a) *to*

$$\theta_{n,G}=\theta_{n-1,G}+(\alpha+\rho_n w_n)r_{n-d}^{-1}(y_n-\hat{y}_{n,G})\phi_{n-d,G}, \tag{3.31}$$

where $\phi_{n,G}$, $\hat{y}_{n,G}$, and r_n are still given by (2.21b)-(2.21d), $0<\alpha<a$, *the w_n are random variables satisfying* (3.5) *for some* $c>0$, *and*

$$\rho_n=(a^{1/2}\alpha^{1/2}-\alpha)/\{r_{n-d}^{-1/2}|\theta'_{n-d,G}\phi_{n-d,G}|\vee c\}. \tag{3.32}$$

The initial value $\theta_{0,G}$ is so chosen that $b_{0,G}\neq 0$, where $b_{n,G}$ denotes the component of $\theta_{n,G}$ estimating b_1. Then (2.22) *and* (2.23) *still hold and*

$$P\{b_{n,G}=0\}=0 \quad \textit{for every } n. \tag{3.33}$$

Moreover, (2.24) *also holds if $B(z)$ is stable.*

4. Occasional excitation and strongly consistent parameter estimates using the method of moments. Throughout this section we will let $n_1 < m_1 < n_2 < m_2 < \cdots$ be stopping times such that (3.4) holds, and define the modified stochastic gradient algorithm $\theta_{n,\tilde{G}}$ and the associated pseudo-regression vector $\phi_{n,\tilde{G}}$ as in Theorem 2. Letting

$$J = \bigcup_{i=1}^{\infty} \{n_i, n_i+1, \cdots, m_i - 1\}, \tag{4.1}$$

we will assume that, as in Theorem 2, the inputs u_n are determined by (3.8) for $n \in J$. Making use of Corollary 1, we show in the following theorem that the input-output data at stages $t \in J$ have certain excitation properties, which we will use later for the construction of strongly consistent estimates of the parameter vector θ, defined by (1.6), of system (1.1).

THEOREM 4. *With the same notation and assumptions as in Theorem 2, assume that $B(z)$ is stable. Suppose that*

$$m_j - n_j \textit{ is } \mathscr{F}_{nj}\textit{-measurable}, \quad m_j - n_j \to \infty, \quad \textit{and}$$

$$\left(1 \vee \sum_{\nu=0}^{p(d)-1} y^2_{n_j-\nu} \vee \sum_{\nu=1}^{k+d-2} u^2_{n_j-\nu}\right) \log^2 \left\{ \sum_{i=1}^{j} \left(2 \vee \sum_{\nu=0}^{p(d)-1} y^2_{n_i-\nu} \vee \sum_{\nu=1}^{k+d-2} u^2_{n_i-\nu} \right)\right\} \tag{4.2}$$

$$= O(m_j - n_j) \quad \textit{a.s.}$$

Let $\#_j = \sum_{i=1}^{j} (m_i - n_i)$ *and define*

$$\begin{aligned} X_t &= (y_{t-1}, \cdots, y_{t-p}, u_{t-d}, \cdots, u_{t-d-k+1})', \\ Z_t &= (y_{t-h-1}, \cdots, y_{t-h-k+1}, w_{t-d}, \cdots, w_{t-d-p-k+1})'. \end{aligned} \tag{4.3}$$

Then (3.24), (3.25) *hold and there exists* $(p+2k-1)\times(p+k)$ *nonrandom matrix H such that*

$$\lim_{j\to\infty} \left(\sum_{i=1}^{j} \sum_{t=n_i+p+k+h+d}^{m_i} Z_t X_t' \right) \Big/ \#_j = H \quad \textit{a.s.} \tag{4.4}$$

If the polynomials $z^p A(z^{-1})$, $z^{k-1} B(z^{-1})$ and $z^h C(z^{-1})$ are coprime, then H has full rank $p+k$.

Proof. In view of (4.2), (3.23) holds and therefore by Corollary 1, (3.24) and (3.25) hold. By Lemma 4(i), there exist constants $C > 0$, $0 < \rho < 1$, and $\alpha_0, \alpha_1, \cdots, \beta_0, \beta_1, \cdots$ such that $|\alpha_t| \vee |\beta_t| \leqq C\rho^t$ and

$$\begin{aligned} &u_{t-d} = \alpha_0 y_t + \cdots + \alpha_{t-n} y_n + \beta_0 \varepsilon_t + \cdots + \beta_{t-n}\varepsilon_n + \Delta_{t,n} \quad \text{for } t \geqq n+d, \\ &\text{where } |\Delta_{t,n}| \leqq C\rho^{t-n} \left\{ \sum_{s=1}^{k-1} |u_{n-d-s}| + \sum_{s=1}^{p} |y_{n-s}| + \sum_{s=1}^{h} |\varepsilon_{n-s}| \right\}. \end{aligned} \tag{4.5}$$

From the definition of Δ_{t,n_i+d}, it follows that

$$\begin{aligned} \sum_{i=1}^{j} \sum_{t=n_1+1+d}^{m_i} \Delta^2_{t,n_i+d} &\leqq (k+p+h) \left(C^2 \sum_{s=1}^{\infty} \rho^{2s} \right) \sum_{i=1}^{j} \left(\sum_{\nu=1}^{k-1} u^2_{n_i-\nu} \right. \\ &\quad \left. + \sum_{\nu=1}^{p} y^2_{n_i+d-\nu} + h \sup_t \varepsilon_t^2 \right) = o(\#_j) \quad \text{a.s. by (4.2) and (3.18).} \end{aligned} \tag{4.6}$$

Let $\delta_t = y_t - \eta_t - w_{t-d}$. By (3.25), $\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i} \delta_t^2 = o(\#_j)$ almost surely and therefore

$$\begin{aligned} \sum_{i=1}^{j} \sum_{t=n_1+1+d}^{m_i} \left(\sum_{s=n_i+d}^{t} |\alpha_{t-s}\delta_s| \right)^2 &\leqq \sum_{i=1}^{j} \sum_{t=n_i+1+d}^{m_i} \left(\sum_{s=n_i+d}^{t} |\alpha_{t-s}| \right) \left(\sum_{s=n_i+d}^{t} |\alpha_{t-s}| \delta_s^2 \right) \\ &\leqq \sum_{i=1}^{j} \sum_{s=n_i+d}^{m_i} \delta_s^2 \left(\sum_{t=0}^{\infty} |\alpha_t| \right)^2 - o(\#_j) \quad \text{a.s.} \end{aligned} \tag{4.7}$$

For fixed $r \geqq 1$ and $s \geqq 1$, $s' \geqq 0$, since ε_{t-r} is $\mathcal{F}_{t-1}$-measurable and $\{n_i + s \leqq t \leqq m_i - s'\} \in \mathcal{F}_{t-s}$ by (4.2). We can apply Lemma 1(i) to conclude that for every $\delta > 0$,

$$\sum_{i=1}^{j} \sum_{t=n_i+s}^{m_i-s'} \varepsilon_{t-r}\varepsilon_t = o(\#_j^{1/2+\delta}) \quad \text{a.s.} \tag{4.8}$$

Moreover, by (2.26),

$$\sum_{i=1}^{j} \sum_{t=n_i+s}^{m_i-s'} \varepsilon_t^2 \sim \sigma^2 \#_j \quad \text{a.s.} \tag{4.9}$$

Likewise, in view of (3.5), we have

$$\sum_{i=1}^{i} \sum_{t=n_i+s}^{m_i-s'} w_t^2 \sim v \#_j \quad \text{a.s.} \tag{4.10}$$

$$\left| \sum_{i=1}^{j} \sum_{t=n_i+s}^{m_i-s'} \varepsilon_r w_t \right| + \left| \sum_{i=1}^{j} \sum_{t=n_i+s}^{m_i-s'} \varepsilon_t w_{t-r} \right| + \left| \sum_{i=1}^{j} \sum_{t=n_i+s}^{m_i-s'} \varepsilon_{t-r} w_t \right| = o(\#_j^{1/2+\delta}) \quad \text{a.s.} \tag{4.11}$$

Since $\eta_t = \varepsilon_t + f_1\varepsilon_{t-1} + \cdots + f_{d-1}\varepsilon_{t-d+1}$, it follows from (4.8)-(4.11) that for every fixed L,

$$\begin{aligned}
\lim_{j\to\infty} \Bigg\{ \sum_{i=1}^{j} \sum_{t=n_i+p+k+h+d}^{m_i} & (\eta_{t-h-1} + w_{t-h-1-d}, \cdots, \eta_{t-h-k+1} \\
& + w_{t-h-k+1-d}, w_{t-d}, \cdots, w_{t-p-k+1-d})' \\
\times \Bigg(& \eta_{t-1} + w_{t-1-d}, \cdots, \eta_{t-p} + w_{t-p-d}, \sum_{r=0}^{L} \alpha_r(\eta_{t-r} + w_{t-r-d}) + \beta_r\varepsilon_{t-r}, \cdots, \\
& \sum_{r=0}^{L} \alpha_r(\eta_{t-k+1-r} + w_{t-k+1-r-d}) + \beta_r\varepsilon_{t-k+1-r} \Bigg) \Bigg\} \Big/ \#_j = H_L \quad \text{a.s.,}
\end{aligned} \tag{4.12}$$

where $H_L = H_L(\alpha_0, \cdots, \alpha_L, \beta_0, \cdots, \beta_L, f_1, \cdots, f_{d-1}, \sigma^2, v)$ is a nonrandom matrix.

Since $\sup_t (|\varepsilon_t| + |w_t|) < \infty$ almost surely, and since $\sum_0^\infty (|\alpha_r| + |\beta_r|) < \infty$,

$$\lim_{L\to\infty} \sup_j \left\{ \sum_{i=1}^{j} \sum_{t=n_i}^{m_i} \sup_{0\leqq s<t} (\varepsilon_{t-s}^2 + w_{t-s}^2) \sum_{r=L+1}^{\infty} (|\alpha_r| + |\beta_r|) \right\} \Big/ \#_j = 0 \quad \text{a.s.} \tag{4.13}$$

Since $y_t = \eta_t + w_{t-d} + \delta_t$ with $\sum_{i=1}^{j} \sum_{t=n_i+d}^{m_i} \delta_t^2 = o(\#_j)$ almost surely, it follows from (4.5)-(4.7) together with (4.12) and (4.13) that (4.4) holds with $H (= \lim_{L\to\infty} H_L)$ equal to the common expected value of the stationary sequence of random matrices

$$\begin{aligned}
(\tilde\eta_{t-h-1} + \tilde w_{t-h-1-d}, \cdots, \tilde\eta_{t-h-k+1} + \tilde w_{t-h-k+1-d}, \tilde w_{t-d}, \cdots, \tilde w_{t-d-p-k+1})' \\
\times \Bigg(\tilde\eta_{t-1} + \tilde w_{t-1-d}, \cdots, \tilde\eta_{t-p} + \tilde w_{t-p-d}, \sum_{j=0}^{\infty} \{\alpha_j(\tilde\eta_{t-j} + \tilde w_{t-j-d}) + \beta_j\tilde\varepsilon_{t-j}\}, \cdots, \\
\sum_{j=0}^{\infty} \{\alpha_j(\tilde\eta_{t+1-k-j} + \tilde w_{t+1-k-j-d}) + \beta_j\tilde\varepsilon_{t+1-k-j}\} \Bigg),
\end{aligned}$$

where $\tilde\eta_t = \tilde\varepsilon_t + \cdots + f_{d-1}\tilde\varepsilon_{t-d+1}$ and $\cdots, \tilde\varepsilon_{-1}, \tilde w_{-1}, \tilde\varepsilon_0, \tilde w_0, \tilde\varepsilon_1, \tilde w_1, \cdots$ are independent, bounded random variables such that the $\tilde\varepsilon_i$ have the same distribution with mean 0 and variance σ^2, and the $\tilde w_i$ have a common distribution with mean 0 and variance v. Let $\tilde u_{t-d} = \sum_{j=0}^{\infty} \{\alpha_j(\tilde\eta_{t-j} + \tilde w_{t-j-d}) + \beta_j\tilde\varepsilon_{t-j}\}$. Then by the definitions of α_j and β_j,

$$A(q^{-1})(\tilde\eta_t + \tilde w_{t-d}) = B(q^{-1})\tilde u_{t-d} + C(q^{-1})\tilde\varepsilon_t.$$

Hence by Lemma 6, H has full rank $p+k$. □

In view of Theorem 4 on occasionally excited input-output systems, we can use the instrumental variables Z_t defined in (4.3) to construct strongly consistent estimates of $a_1, \cdots, a_p, b_1, \cdots, b_k$ based on the input-output data at stages $t \in J$. We can then estimate consistently the autocorrelation function of the colored-noise sequence $\{\varepsilon_n + c_1\varepsilon_{n-1} + \cdots + c_h\varepsilon_{n-h}\}$.

COROLLARY 2. *With the same notation and assumptions as in Theorem* 4, *suppose that the polynomials* $z^pA(z^{-1})$, $z^{k-1}B(z^{-1})$, *and* $z^hC(z^{-1})$ *are coprime. Let* $\lambda = (-a_1, \cdots, -a_p, b_1, \cdots, b_k)'$. *Define*

$$\lambda_{m_j} = (V_j'V_j)^{-1}V_j' \sum_{i=1}^{j} \sum_{t=n_i+p+k+h+d}^{m_i} Z_t y_t \quad \textit{where}$$
$$V_j = \sum_{i=1}^{j} \sum_{t=n_i+p+k+h+d}^{m_i} Z_t X_t', \tag{4.14}$$

and the inverse is the Moore–Penrose generalized inverse. Then

$$\lambda_{m_j} - \lambda = o(\#_j^{-1/2+\delta}) \quad a.s. \tag{4.15}$$

for every $\delta > 0$. *Let*

$$e_{j,t} = y_t - \lambda_{m_j}' X_t, \tag{4.16}$$

$$\rho_{m_j}(\nu) = \left(\sum_{i=1}^{j} \sum_{t=n_i+p+k+h+d}^{m_i} e_{j,t}e_{j,t-\nu} \right) \Big/ \left(\sum_{i=1}^{j} \sum_{t=n_i+p+k+h+d}^{m_i} \right), \quad \nu = 0, \cdots, h. \tag{4.17}$$

Then $\rho_{m_j}(\nu)$ *converges almost surely to the covariance* $\rho(\nu)$ *between* $C(q^{-1})\varepsilon_t$ *and* $C(q^{-1})\epsilon_{t-\nu}$ (*which is the same for all t since* $E(\varepsilon_t^2 | \mathscr{F}_{t-1}) = \sigma^2$); *in fact, for every* $\delta > 0$,

$$\rho_{m_j}(\nu) - \rho(\nu) = o(\#_j^{-1/2+\delta}) \quad \textit{a.s. for } \nu = 0, 1, \cdots, h. \tag{4.18}$$

Remark. In view of (4.16), the numerator of (4.17) can be expressed as

$$\begin{aligned}\sum_{i=1}^{j} \sum_{t=n_i+p+k+h+d}^{m_i} e_{j,t}e_{j,t-\nu} = {} & \sum_{i=1}^{j} \sum_{t=n_i+p+k+h+d}^{m_i} y_t y_{t-\nu} \\ & - \lambda_{m_j}' \sum_{i=1}^{j} \sum_{t=n_i+p+h+k+d}^{m_i} (X_{t+\nu}y_t + X_t y_{t-\nu}) \\ & + \lambda_{m_j}' \left(\sum_{i=1}^{j} \sum_{t=n_i+p+k+h+d}^{m_i} X_t X_{t-\nu}' \right) \lambda_{m_j},\end{aligned}$$

which can therefore be updated at stage m_j without calculating the residuals $e_{j,t}$.

Proof. Since $y_t = X_t'\lambda + \varepsilon_t + c_1\varepsilon_{t-1} + \cdots + c_h\varepsilon_{t-h}$, it follows from (4.14) and (4.3) that

$$\begin{aligned}\lambda_{m_j} = \lambda + (V_j'V_j)^{-1}V_j \sum_{i=1}^{j} \sum_{t=n_i+p+k+h+d}^{m_i} & \\ (\varepsilon_t + \cdots + c_h\varepsilon_{t-h})(y_{t-h-1}, \cdots, y_{t-h-k+1}, w_{t-d}, \cdots, w_{t-d-p-k+1})'.&\end{aligned} \tag{4.19}$$

In view of (3.5), (3.24), and Lemma 1(i), we have for every $\delta > 0$,

$$\begin{aligned}\sum_{i=1}^{j} \sum_{t=n_i+p+k+h}^{m_i} (\varepsilon_t + \cdots + c_h\varepsilon_{t-h})(y_{t-h-1}, \cdots, & \\ y_{t-h-k+1}, w_{t-d}, \cdots, w_{t-d-p-k+1})' = o(\#_j^{1/2+\delta}) \quad \text{a.s.}&\end{aligned} \tag{4.20}$$

Since $V_j / \#_j \to H$ almost surely and since $H'H$ is positive definite by Theorem 4, (4.15) follows from (4.19) and (4.20).

Noting that $e_{j,t}=C(q^{-1})\varepsilon_t+(\lambda-\lambda_{m_j})'X_t$ and that $\sum_{i=1}^{j}\sum_{t=n_i+p+k+h+d}^{m_i}(\|X_t\|+\|X_t\|^2)=O(\#_j)$ almost surely by an argument similar to the proof of Theorem 4, we obtain from (4.15) that for every $\delta>0$,

$$\sum_{i=1}^{j}\sum_{t=n_i+p+k+h+d}^{m_i}\{e_{m_j,t}\,e_{m_j,t-\nu}-[C(q^{-1})\varepsilon_t][C(q^{-1})\varepsilon_{t-\nu}]\}=o(\#_j^{1/2+\delta})\quad\text{a.s.}\tag{4.21}$$

for $\nu=0,1,\cdots,h$. Moreover, by (4.9) and (4.10),

$$\begin{aligned}\sum_{i=1}^{j}\sum_{t=n_i+p+k+h+d}^{m_i}\{(\varepsilon_t+c_1\varepsilon_{t-1}+\cdots+c_h\varepsilon_{t-h})(\varepsilon_{t-\nu}+\cdots+c_h\varepsilon_{t-\nu-h})-\rho(\nu)\}\\=o(\#_j^{1/2+\delta})\quad\text{a.s.}\end{aligned}\tag{4.22}$$

for every $\delta>0$. From (4.21) and (4.22), (4.18) follows. □

Suppose that m_j is so chosen that in addition to (4.2) we also have

$$m_j-n_j=O(\#_{j-1}),\ \ \#_{j-1}=O(m_j-n_j)\quad\text{a.s.}\tag{4.23}$$

Define an estimate $\mathbf{s}_j=(\sigma_0^{(j)},\cdots,\sigma_h^{(j)})'$ of $\mathbf{s}=(\sigma,c_1\sigma,\cdots,c_h\sigma)'$ by the following one-step modification of Wilson's [22] iterative algorithm

$$\mathbf{s}_{j+1}=\mathbf{T}_j^{-1}(f_j(0)+\rho_{m_j}(0),\cdots,f_j(h)+\rho_{m_j}(h))',\tag{4.24}$$

where $f_j(\nu)=\sum_{i=0}^{h-\nu}\sigma_i^{(j)}\sigma_{i+\nu}^{(j)}(\nu=0,\cdots,h)$ and $\mathbf{T}_j$ is a $(h+1)\times(h+1)$-matrix whose (μ,ν) element is $\sigma_{\mu+\nu}^{(j)}+\sigma_{\nu-\mu}^{(j)}$, setting $\sigma_i^{(j)}=0$ for $i>h$ or $i<0$. The basic idea behind (4.24) is essentially the same as Solo's [23] recursive implementation of Wilson's algorithm, and analogous to [23], we initialize (4.24) by setting $\sigma_0^{(0)}>0=\sigma_1^{(0)}=\cdots=\sigma_h^{(0)}$ and perform a stability test for the polynomial $\sum_{\nu=0}^{h}\sigma_\nu^{(j+1)}z^\nu$, redefining (4.24) by $\mathbf{s}_{j+1}=\mathbf{s}_j$ if the stability test fails.

From (4.23) it follows that $\limsup_{j\to\infty}\#_{j-1}/\#_j<1$ almost surely and therefore for every $\delta>0$, $\sum_{j=1}^{\infty}\#_j^{-1/2+\delta}<\infty$ almost surely, implying that

$$\prod_{j=1}^{n}(1+\#_j^{-1/2+\delta})\ \text{converges a.s. as}\ n\to\infty.\tag{4.25}$$

Furthermore, (4.23) implies that $\#_j=O(\#_{j-1})$ almost surely, and by (4.18), for every $\delta>0$,

$$\max_{0\le\nu\le h}|\rho_{m_j}(\nu)-\rho_{m_{j-1}}(\nu)|=O(\#_j^{-1/2+\delta})\quad\text{a.s.}\tag{4.26}$$

Hence a straightforward modification of Solo's argument in [23] can be used to show that $\mathbf{s}_j\to\mathbf{s}$ almost surely; moreover, further application of (4.26) gives the convergence rate $\mathbf{s}_j-\mathbf{s}=O(\#_j^{-1/2+\delta})$ almost surely for every $\delta>0$ (cf. [24]). Throughout the sequel we will let

$$\theta_{m_j}=(\lambda'_{m_j},\sigma_1^{(j)}/\sigma_0^{(j)},\cdots,\sigma_h^{(j)}/\sigma_0^{(j)})'.\tag{4.27}$$

Therefore under the assumptions of Corollary 2 and (4.23), θ_{m_j} provides a strongly consistent estimator of $(\lambda',c_1,\cdots,c_h)'(=\theta$, by (1.6)); in fact,

$$\theta_{m_j}=\theta+o(\#_j^{-1/2+\delta})\quad\text{a.s. for every}\ \delta>0.\tag{4.28}$$

5. Matrix gain and the monitored recursive maximum likelihood algorithm. By introducing occasional white-noise perturbations and applying the modified stochastic gradient algorithm $\theta_{n,\tilde{G}}$ in conjunction with these white-noise perturbations to define the inputs u_n for $n\in J$ as in (3.8), we showed in § 4 that the input-output data at stages

$t \in J$ have sufficient excitation properties to provide strongly consistent estimators θ_{m_j} of θ. These estimators are updated only occasionally at stages $n = m_j$ and therefore can afford more computational complexities than the stochastic gradient algorithms $\theta_{n,G}$, $\theta_{n,\tilde{G}}$, and the monitored recursive maximum likelihood algorithm $\theta_{n,M}$, to be introduced below, which must be updated at every stage n.

The strongly consistent estimate θ_{m_j} of the parameter vector θ in the explicit dynamical system (1.1) induces a strongly consistent estimate $\tilde{\theta}_{m_j}$ of the reparametrization $\tilde{\theta}$ in the implicit system (2.17), since the components of $\tilde{\theta}$ can be expressed as continuous functions of those of θ in view of (2.12). For example, in the case $d = 2$, it follows from (2.12) that $f_1 = c_1 - a_1$, $g_i = c_{i+1} - a_{i+1} - a_i f_1$ for $i = 1, \cdots, p \vee (h-1)$, where we set $c_i = 0$ for $i \geqq h$ and $a_j = 0$ for $j > p$; moreover, $(bf)_i = b_i + b_{i-1} f_1$ for $i = 2, \cdots, k$, and $(bf)_{k+1} = b_k f_1$. The convergence rate (4.28) for θ_{m_j} implies that in the reparametrized model (2.17) we again have

$$\tilde{\theta}_{m_j} = \tilde{\theta} + o(\#_j^{-1/2+\delta}) \quad \text{a.s. for every } \delta > 0. \tag{5.1}$$

Therefore, taking $0 < \delta < 1/2$ and $c > 0$ and letting $\mathbf{I}_j$ denote the cube in $R^{p(d)+k+d-1+h}$ with center at $\tilde{\theta}_{m_j}$ and width $c\#_j^{-1/2+\delta}$, we obtain from (5.1) that

$$P\{\tilde{\theta} \in \mathbf{I}_j \text{ for all large } j\} = 1; \tag{5.2}$$

moreover, $\mathbf{I}_j$ shrinks to $\tilde{\theta}$ as $j \to \infty$ with probability 1.

The basic ideas of the monitored recursive maximum likelihood algorithm $\theta_{n,M}$ for the estimation of $\tilde{\theta}$ are: (i) to extend the classical recursive maximum likelihood (RML2; cf. [25]) algorithm to the implicit system (2.17), and (ii) to constrain (monitor) the recursive algorithm so that it lies inside the "confidence interval" $\mathbf{I}_j$ for $\nu_j \leqq n < \nu_{j+1}$, where $(\nu_{j+1} >) \nu_j \geqq m_j$ is so chosen to accommodate the time for carrying out the more intensive computations required by $\tilde{\theta}_{m_j}$. The projection that we use to constrain $\theta_{n,M}$ is taken with respect to the norm induced by the positive-definite matrix P_{n-d}^{-1} defined in (5.4d) below, instead of the usual Euclidean norm. For $x \in R^{p(d)+k+d-1+h}$ and $\nu_j \leqq n < \nu_{j+1}$, let $\pi(x)$ denote the unique solution of the quadratic programming problem

$$(\pi_n(x) - x)' P_{n-d}^{-1} (\pi_n(x) - x) = \min_{y \in \mathbf{I}_j} \{(y-x)' P_{n-d}^{-1} (y-x)\}, \tag{5.3}$$

i.e., $\pi_n(x)$ is the projection of x into $\mathbf{I}_j$ with respect to the norm induced by P_{n-d}^{-1}. The choice of a cube for the confidence region $\mathbf{I}_j$ implies linear constraints for the quadratic programming problem (5.3), which can be handled by simple computational methods (cf. [26]). Define

$$\theta_{n,M} = (\hat{g}_{n,1}, \cdots, \hat{g}_{n,p(d)}, \hat{b}_{n,1}, (\widehat{bf})_{n,2}, \cdots, (\widehat{bf})_{n,k+d-1}, -\hat{c}_{n,1}, \cdots, -\hat{c}_{n,h})'$$

for $n > \nu_1$ by the recursion

$$\theta_{n,M} = \pi_n(\theta_{n-1,M} + P_{n-d} \xi_{n-d} (y_n - \hat{y}_{n,M})), \tag{5.4a}$$

$$\begin{gathered} \xi_n + \hat{c}_{n,1} \xi_{n-1} + \cdots + \hat{c}_{n,h} \xi_{n-h} = \phi_{n,M}, \quad \text{where} \\ \phi_{n,M} = (y_n, \cdots, y_{n-p(d)+1}, u_n, \cdots, u_{n-k-d+2}, \hat{y}_{n+d-1,M}, \cdots \hat{y}_{n+d-h,M})', \end{gathered} \tag{5.4b}$$

$$\hat{y}_{n+d,M} = \theta'_{n,M} \phi_{n,M}, \tag{5.4c}$$

$$P_n^{-1} = P_{n-1}^{-1} + \xi_n \xi'_n + I/n, \tag{5.4d}$$

where P_{ν_1} is a positive-definite matrix and $\theta_{\nu_1,M}$ represents an initial guess of $\tilde{\theta}$ (e.g., set $\theta_{\nu_1,M} = \tilde{\theta}_{m_1}$), and I denotes the identity matrix.

The recursive algorithm $\theta_{n,M}$ in (5.4) involves the $\mathcal{F}_{n-d}$-measurable matrix gain $P_{n-d}=\{P_{n-d-1}^{-1}+\xi_{n-d}\xi'_{n-d}+I/(n-d)\}^{-1}$. Since $\pi_n(x)$ is the projection of x into the closed convex set $\mathbf{I}_j$ with respect to the norm induced by P_{n-d}^{-1}, it follows from (5.4a) that the extended stochastic Lyapunov function $Q_n \triangleq (\theta_{n,M}-\tilde{\theta})'P_{n-d}^{-1}(\theta_{n,M}-\tilde{\theta})$ satisfies

$$\begin{aligned} Q_n &\leqq (\theta_{n-1,M}-\tilde{\theta}+P_{n-d}\xi_{n-d}(y_n-\hat{y}_{n,M}))'P_{n-d}^{-1}(\theta_{n-1,M}-\tilde{\theta}+P_{n-d}\xi_{n-d}(y_n-\hat{y}_{n,M})) \\ &= Q_{n-1}+(n-d)^{-1}\|\theta_{n-1,M}-\tilde{\theta}\|^2+[\xi'_{n-d}(\theta_{n-1,M}-\tilde{\theta})]^2 \\ &\qquad +2\xi'_{n-d}(\theta_{n-1,M}-\tilde{\theta})(y_n-\hat{y}_{n,M})+\xi'_{n-d}P_{n-d}\xi_{n-d}(y_n-\hat{y}_{n,M})^2, \end{aligned}$$

for $\nu_j \leqq n < \nu_{j+1}$. By making use of Lemmas 1, 2, and 4(ii) to analyze the above recursive inequalities, we have established in [16] the following basic property for the adaptive d-step ahead predictors $\hat{y}_{n+d,M}=\theta'_{n,M}\phi_{n,M}$, which will play a key role in the development in § 6 of asymptotically efficient certainty-equivalence-type rules whose regrets have the logarithmic order (1.24) or (1.25).

THEOREM 5. *Suppose that $C(z)$ is stable and that the random disturbances ε_n in the linear stochastic system* (1.1) *satisfy assumption* (1.3) *and the boundedness condition* $\sup_n |\varepsilon_n|<\infty$ *almost surely. Let $\nu_1<\nu_2<\cdots$ be stopping times with respect to $\{\mathcal{F}_t\}$ and let $\mathbf{I}_j$ be an $\mathcal{F}_{\nu_j}$-measurable, closed, and convex set such that* (5.2) *holds and*

$$\lim_{j\to\infty} (\textit{diameter of } \mathbf{I}_j)=0 \quad \textit{a.s.} \tag{5.5}$$

Define the monitored recursive maximum likelihood algorithm $\theta_{n,M}$ by (5.4), *where π_n is given by* (5.3) *for $\nu_j \leqq n<\nu_{j+1}$. Assume that*

$$\lambda_{\max}(P_n^{-1})\to\infty \textit{ and } \xi'_nP_n\xi_n\to 0 \quad \textit{a.s.} \tag{5.6}$$

Let $\sigma_d^2=E\eta_d^2$, where η_t is given in (2.15). *Then*

$$\begin{aligned} \sum_{i=1}^{n} \{E(y_{t+d}\,|\,\mathcal{F}_t)-\hat{y}_{t+d,M}\}^2 &\leqq (2d-1)(\sigma_d^2+o(1))(\log\det P_n^{-1}) \\ &\quad +o(\log n)+o\left(\sum_{j:\nu_j\leqq n+d}\ \sum_{r=1}^{h+d-1}[\|\xi_{\nu_j-r}\|^2+\|\xi_{\nu_j-r}\|]\right) \quad \text{a.s.} \end{aligned} \tag{5.7}$$

6. Adaptive control schemes based on parallel recursive identification algorithms. In this section we construct a class of certainty-equivalence-type rules that involve parallel implementation of the recursive identification algorithms $\theta_{n,G}$, $\theta_{n,\tilde{G}}$, and $\theta_{n,M}$ described in §§ 3 and 5 and whose control inputs are determined by one of the following three equations at every stage t: (i) $\theta'_{t,G}\phi_{t,G}=y^*_{t+d}$, (ii) $\theta'_{t,M}\phi_{t,M}=y^*_{t+d}$, (iii) $\theta'_{t,\tilde{G}}\phi_{t,\tilde{G}}=w_t$, where y^*_{t+d} represents the target value at stage $t+d$, and the w_t represent extraneous white-noise perturbations that satisfy condition (3.5). We will show that such adaptive control schemes have logarithmic order (1.24) for the regret $\sum_{d+1}^{n}(y_i-y_i^*-\eta_i)^2$ under the following assumptions on the linear stochastic system (1.1):

(6.1) $B(z)$ is stable.

(6.2) $\operatorname{Re}\{C(e^{it})-(d-1/2)a\}>0$ for all $|t|\leqq\pi$ and some $a>0$.

(6.3) The polynomials $z^pA(z^{-1})$, $z^{k-1}B(z^{-1})$ and $z^hC(z^{-1})$ are coprime.

(6.4) $\{\varepsilon_n\}$ satisfies (1.3) and $\sup_n|\varepsilon_n|<\infty$ almost surely.

(6.5) The target value y_n^* at stage n is $\mathcal{F}_{n-d}$-measurable and $\sum_1^n y_i^{*2}=O(n)$ almost surely, $y_n^{*2}=o((\log n)^\gamma)$ almost surely for some $0<\gamma<1$.

For the stochastic gradient algorithm $\phi_{n,G}$, we use the version defined by (3.31), (3.32), and (2.21b)-(2.21d). To define the sequence of stopping times $n_1 < m_1 < n_2 < \cdots$ associated with the modified stochastic gradient algorithm $\theta_{n,\tilde{G}}$ (cf. Theorem 2), take a nondecreasing sequence of constants $K_n \geqq d+h$ such that

$$K_n \to \infty,\ K_n = O((\log n)^{(1/2)\wedge(1-\gamma)}),\ K_{2n} = O(K_n), \tag{6.6}$$

where $\gamma(<1)$ is the same as in (6.5). Define inductively

$$\begin{aligned} n_j = \inf\Bigg\{ n > m_{j-1}\colon \sum_{i=1}^{j-1} (m_i - n_i) \leqq K_n \text{ and} \\ \left(\sum_{r=0}^{p(d)-1} y_{n-r}^2\right) \vee \left(\sum_{r=1}^{k+d-2} u_{n-r}^2\right) \leqq K_n/\log^2(K_n+2)\Bigg\}, \end{aligned} \tag{6.7}$$

$$\begin{aligned} m_j = n_j + \Bigg[K_{n_j} \vee \left(1 \vee \sum_{r=0}^{p(d)-1} y_{n_j-r}^2 \vee \sum_{r=1}^{k+d-2} u_{n_j-r}^2\right) \\ \times \log^2\left\{\sum_{i=1}^{j}\left(2 \vee \sum_{r=0}^{p(d)-1} y_{n_i-r}^2 \vee \sum_{r=1}^{k+d-2} u_{n_i-r}^2\right)\right\}\Bigg]. \end{aligned} \tag{6.8}$$

Note that the n_j and m_j defined by induction in this way are stopping times with respect to $\{\mathcal{F}_t\}$ and satisfy conditions (3.4) and (4.2).

Let $J = \bigcup_i \{n\colon n_i \leqq n \leqq m_i - 1\}$, as in (4.1), $J_n = \{t \in J\colon t \leqq n\}$, and use (3.8) to define the input u_n at stage n if $n \in J$. Using $\#(S)$ to denote the number of elements of a set S, define u_n for $n \notin J$ by

$$\theta'_{n,G}\phi_{n,G} = y^*_{n+d}, \text{ if } \#(J_n) \leqq K_n^{1/2} \quad \text{and } n \notin J. \tag{6.9}$$

By (3.33), (6.9) is well defined almost surely. Note that n_j is the first time n after m_{j-1} for which $\#(J_{n-1}) \leqq K_n$ (signalling too few white-noise excitations) and the subvectors $(y_n, \cdots, y_{n-p(d)+1})'$ and $(u_{n-1}, \cdots, u_{n-k-d+2})'$ of $\phi_{n,G}$ have squared lengths $\leqq K_n/\log^2(K_n+2)$, and that $m_j - n_j$ is a simple function of K_{n_j} and these components of $\phi_{n,G}$. By Lemma 7 below, the stopping times n_j and m_j defined by (6.7) and (6.8) are indeed finite almost surely. Moreover, they satisfy condition (4.23) in view of Lemma 9(i) below.

With n_j and m_j thus defined, define the recursive method-of-moments estimator θ_{m_j} as in § 4, which induces a strongly consistent estimate $\tilde{\theta}_{m_j}$ of the reparametrization $\tilde{\theta}$ in the implicit system (2.17). The cube $\mathbf{I}_j$ with center at $\tilde{\theta}_{m_j}$ and width $c\#_j^{-1/2+\delta}$ is used to define the monitored recursive maximum likelihood algorithm $\theta_{n,M} = (\hat{g}_{n,1}, \cdots, \hat{g}_{n,p(d)}, \hat{b}_{n,1}, \cdots, -\hat{c}_{n,h})'$ by (5.4), in which π_n is given by (5.3) for $\nu_j \leqq n \leqq \nu_{j+1}$, where $\nu_j(\geqq m_j)$ represents the time at which the auxiliary consistent estimate $\tilde{\theta}_{m_j}$ is available for monitoring the $\theta_{n,M}$. *In conjunction with* (3.8) *and* (6.9), we complete the specification of the input sequence $\{u_n\}$ by

$$\theta'_{n,M}\phi_{n,M} = y^*_{n+d} \quad \text{if } \#(J_n) > K_n^{1/2} \text{ and } n \notin J\ (n > \nu_1), \tag{6.10}$$

setting $u_n = w_n$ in the case where $\hat{b}_{n,1} = 0$ (for which u_n is not well defined by (6.10)), and applying (6.9) when $n \leqq \nu_1$. Since $\theta_{n,M} \to \tilde{\theta}$ almost surely and since $b_1 \neq 0$, it follows that with probability 1, $\hat{b}_{n,1} \neq 0$ and u_n is well defined by (6.10) for all large n.

THEOREM 6. *For the sequence of control inputs u_n defined above by* (3.8) *for $n \in J$ and by* (6.9) *and* (6.10) *for $n \notin J$,*

$$\limsup_{n\to\infty} \sum_{i=d+1}^{n} (y_i - y_i^* - \eta_i)^2/\log n \leqq (2d-1)(p(d)+k+d-1+h)E\eta_d^2 \quad \textit{a.s.}, \tag{6.11}$$

under assumptions (6.1)-(6.5), *where* η_i *is defined in* (2.15) *and* $p(d)=p\vee(h-d+1)$. *If furthermore*

$$\log\left(1+\sum_{i=1}^{n} y_i^{*2}\right)=o(\log n)\quad a.s., \tag{6.12}$$

then (6.11) *can be strengthened into*

$$\limsup_{n\to\infty}\sum_{i=d+1}^{n}(y_i-y_i^*-\eta_i)^2/\log n\leqq(2d-1)\{p(d)+k+d-2\}E\eta_d^2\quad a.s. \tag{6.13}$$

We preface the proof of Theorem 6 by the following five lemmas.

LEMMA 7. *Suppose that the inputs* u_n *are so chosen that* (6.9) *holds. Then* $\#(J)=\infty$ *almost surely.*

Proof. On the event $\{\#(J)<\infty\}$, it follows from (6.9) that $\theta'_{n,G}\phi_{n,G}=y^*_{n+d}$ for all large n, and therefore by (2.24) (cf. Theorem 3),

$$\sum_{n=1}^{N}\left\{\sum_{r=0}^{p(d)-1}y_{n-r}^2+\sum_{r=1}^{k+d-2}u_{n-r}^2\right\}=O(N)\quad \text{a.s. on } \{\#(J)<\infty\}, \tag{6.14}$$

noting that $\sum_1^N(y_i^*)^2=O(N)$ almost surely by (6.5). Since $K_n\to\infty$, (6.14) implies that

$$P\{\#(J)<\infty\}=P\left\{\#(J)<\infty \text{ and } \sum_{r=0}^{p(d)-1}y_{n-r}^2+\sum_{r=1}^{k+d-2}u_{n-r}^2\right.$$
$$\left.\leqq K_n/\log^2(K_n+2) \text{ for infinitely many } n\text{'s}\right\}=0,$$

where the last equality follows from (6.7). Hence $\#(J)=\infty$ almost surely. □

Define the following subsets of integers $\geqq d+1$:

$$\begin{aligned}\Lambda_N&=\{t\leqq N: t-d\notin J \text{ and } \#(J_{t-d})\leqq K_{t-d}^{1/2}\}=\{t\leqq N: (6.9) \text{ is used at } n=t-d\},\\ L_N&=\{t\leqq N: t-d\notin J \text{ and } \#(J_{t-d})>K_{t-d}^{1/2}\}=\{t\leqq N: (6.10) \text{ is used at } n=t-d\}.\end{aligned} \tag{6.15}$$

LEMMA 8. (i) $\#(J_n)=O(K_n)$ *almost surely.*

(ii) $\sum_{d+1}^n(y_i-y_i^*-\eta_i)^2=o(n)$ *and* $\sum_1^n(y_i^2+u_i^2)=O(n)$ *almost surely.*

Proof. (i) Let $\#_j=\sum_{i=1}^j(m_i-n_i)$. From (6.7) it follows that

$$\#_{j-1}\leqq K_{n_j},\quad\left(\sum_{r=0}^{p(d)-1}y_{n_j-r}^2\right)\vee\left(\sum_{r=1}^{k+d-2}u_{n_j-r}^2\right)\leqq K_{n_j}/\log^2(K_{n_j}+2). \tag{6.16}$$

Clearly (6.8) implies that

$$\sum_{i=1}^{j}\left(2\vee\sum_{r=0}^{p(d)-1}y_{n_i-r}^2\vee\sum_{r=1}^{k+d-2}u_{n_i-r}^2\right)=O\left(\sum_{i=1}^{j}(m_i-n_i)\right)=O(\#_j). \tag{6.17}$$

From (6.8), (6.16), and (6.17), it follows that with probability 1, for all large j,

$$\#_j\leqq(m_j-n_j)+K_{n_j}\leqq 2K_{n_j}+\{K_{n_j}/\log^2(K_{n_j}+2)\}\log^2\#_j. \tag{6.18}$$

Since (6.18) implies that $\log\#_j\leqq(1+o(1))\log K_{n_j}$ almost surely, we obtain from (6.18) that $\#_j\leqq(3+o(1))K_{n_j}$ almost surely, giving the desired conclusion in view of the monotonicity of K_n.

(ii) Since $y_t^* = \hat{y}_{t,G}$ for $t \in \Lambda_n$ and $y_t^* = \hat{y}_{t,M}$ for $t \in L_n$, it follows from Corollary 1, (6.18), (6.5), and (6.6) that

$$\sum_1^n (y_t - y_t^* - \eta_t)^2 = \sum_{t \in \Lambda_n} (y_t - \hat{y}_{t,G} - \eta_t)^2 + \sum_{t \in L_n} (y_t - \hat{y}_{t,M} - \eta_t)^2 + o(K_n (\log n)^\gamma) \quad \text{a.s.} \tag{6.19}$$

By Theorem 3, (2.23) holds and therefore by the Kronecker lemma,

$$\sum_1^n (y_t - \hat{y}_{t,G} - \eta_t)^2 = o\left(\sum_1^{n-d} (y_t^2 + u_t^2) + \sum_1^{n-1} \hat{y}_{t,G}^2 \right) + O(1) \quad \text{a.s.} \tag{6.20}$$

Since $C(q^{-1})(y_t - \hat{y}_{t,M} - \eta_t) = -(\theta_{t-d,M} - \tilde{\theta})'\phi_{t-d,M}$ by Lemma 2 and since $\theta_{t-d,M} - \tilde{\theta} \to 0$ almost surely, it follows from the stability of $C(z)$ and Lemma 4(ii) that

$$\begin{aligned} \sum_1^n (y_t - \hat{y}_{t,M} - \eta_t)^2 &= o\left(\sum_1^n \|\phi_{t-d,M}\|^2 \right) + O(1) \\ &= o\left(\sum_1^{n-d} (y_t^2 + u_t^2) + \sum_1^{n-1} \hat{y}_{t,M}^2 \right) + O(1) \quad \text{a.s.} \end{aligned} \tag{6.21}$$

By (6.19)-(6.21) together with (6.4) and (6.5),

$$\begin{aligned} \sum_1^n y_t^2 &\leqq 2 \sum_1^n (y_t^* + \eta_t)^2 + 2 \sum_1^n (y_t - y_t^* - \eta_t)^2 \\ &= O(n) + o\left(\sum_1^{n-d} y_t^2 + \sum_1^{n-d} u_t^2 + \sum_1^{n-1} (\hat{y}_{t,G}^2 + \hat{y}_{t,M}^2) \right) \quad \text{a.s.} \end{aligned} \tag{6.22}$$

Moreover, by (6.20)-(6.22) and (6.4),

$$\begin{aligned} \sum_1^n \hat{y}_{t,G}^2 + \sum_1^n \hat{y}_{t,M}^2 &\leqq 2 \sum_1^n (y_t - \hat{y}_{t,G} - \eta_t)^2 + 2 \sum_1^n (y_t - \hat{y}_{t,M} - \eta_t)^2 \\ &\quad + 4 \sum_1^n (y_t - \eta_t)^2 \\ &= O(n) + o\left(\sum_1^{n-d} y_t^2 + \sum_1^{n-d} u_t^2 + \sum_1^{n-1} (\hat{y}_{t,G}^2 + \hat{y}_{t,M}^2) \right) \quad \text{a.s.} \end{aligned} \tag{6.23}$$

Since $B(z)$ is stable, we can use Lemma 4(i) together with (6.4) to obtain that

$$\sum_1^{n-d} u_t^2 = O\left(\sum_1^n y_t^2 \right) + O(n) \quad \text{a.s.} \tag{6.24}$$

From (6.22)-(6.24), it follows that

$$\sum_1^n (y_t^2 + \hat{y}_{t,G}^2 + \hat{y}_{t,M}^2) = O(n) \quad \text{a.s.}, \qquad \sum_1^{n-d} u_t^2 = O(n) \quad \text{a.s.} \tag{6.25}$$

Combining (6.19)-(6.21) with (6.35), we obtain that $\sum_1^n (y_t - y_t^* - \eta_t)^2 = o(n)$ almost surely. □

LEMMA 9. (i) $\liminf_{n\to\infty} \#(J_n)/K_n > 0$ *almost surely.*

(ii) $\lim_{n\to\infty} \#(\Lambda_n) < \infty$ *almost surely, and therefore* $P\{(6.9)$ *is not used to determine the input* u_n *for all large* $n\} = 1$.

Proof. By Lemma 8(ii), $P(\Omega_0)=1$, where $\Omega_0=\{\limsup_{n\to\infty} n^{-1}\sum_1^n (y_i^2+u_i^2)<\infty\}$. On Ω_0, for all large n,

$$\min_{[n/4]\leqq t\leqq [n/2]}\left\{\sum_{r=0}^{p(d)-1} y_{t-r}^2+\sum_{r=1}^{k+d-2} u_{t-r}^2\right\}<\min_{t\geqq[n/4]} K_t/\log^2(K_t+2). \tag{6.26}$$

In view of the definition of n_j in (6.7), (6.26) implies that either $\#(J_{[n/4]})\geqq K_{[n/4]}$ or there exists $n_j\in\{[n/4],[n/4]+1,\cdots,[n/2]\}$. Since $m_j-n_j\geqq K_{n_j}$ and since K_n is nondecreasing, it then follows that on Ω_0, $\#(J_n)\geqq K_{[n/4]}$ for all large n, and therefore by (6.6), $K_n=O(K_{[n/4]})=O(\#(J_n))$, so there are only finitely many m's for which $\#(J_m)\leqq K_m^{1/2}$. □

LEMMA 10. *$\|\phi_{n,M}\|^2=o(K_n\vee(\log n)^\gamma)$, where γ is given in* (6.5).

Proof. From (3.25), it follows that

$$\max_{n_j+d\leqq t\leqq m_j-1+d}(y_t-w_{t-d}-\eta_t)^2=o(\#_j)\quad\text{a.s.} \tag{6.27}$$

Since $\#_j=O(K_{n_j})$ almost surely by Lemma 8(i) and since $|w_t|\leqq c$, it follows from (6.27) that

$$\lim_{j\to\infty}\left\{\max_{n_j+d\leqq t\leqq m_j-1+d}(y_t-\eta_t)^2/K_t\right\}=0\quad\text{a.s.} \tag{6.28}$$

Defining $\tilde\psi_n$ by (2.17), we obtain by (2.15) that

$$\begin{aligned}\phi_{n,M}-\tilde\psi_n=-(0,\cdots,0,y_{n+d-1}-\hat y_{n+d-1,M}-\eta_{n+d-1},\cdots,&\\ y_{n+d-h}-\hat y_{n+d-h,M}-\eta_{n+d-h})'.&\end{aligned} \tag{6.29}$$

Since by Lemma 2,

$$\begin{aligned}C(q^{-1})(y_n-\hat y_{n,M}-\eta_n)&=-\phi'_{n-d,M}(\theta_{n-d,M}-\tilde\theta)\\ &=-(\phi_{n-d,M}-\tilde\psi_{n-d})'(\theta_{n-d,M}-\tilde\theta)-\tilde\psi'_{n-d}(\theta_{n-d,M}-\tilde\theta),\end{aligned}$$

and since $\theta_{t,M}=(\hat g_{t,1},\cdots,-\hat c_{t,1},\cdots,-\hat c_{t,h})'$, it follows from (6.29) and (2.16) that

$$\hat C_{n-d}(q^{-1})(y_n-\hat y_{n,M}-\eta_n)=-\tilde\psi'_{n-d}(\theta_{n-d,M}-\tilde\theta), \tag{6.30}$$

where $\hat C_t(z^{-1})=1+\hat c_{t,1}z+\cdots+\hat c_{t,h}z^h\to C(z)$ almost surely (as $t\to\infty$). In view of (6.30) and Lemma 4(ii), there exist $D>0$ and $0<\rho_0<1$ such that with probability 1,

$$\begin{aligned}|y_t-\hat y_{t,M}-\eta_t|\leqq D\sup_{i\geqq\nu}\|\theta_{i-d,M}&\\ -\tilde\theta\|\left\{\sum_{i=\nu-d-(p(d)\vee h)}^{t-1}\rho_0^{t-i}(|y_i|+|\eta_i|)+\sum_{i=\nu-k-2d}^{t-d}\rho_0^{t-i}|u_i|\right\}&\\ +D\rho_0^{t-\nu}\sum_{r=0}^{h-1}|y_{\nu-r}-\hat y_{\nu-r,M}-\eta_{\nu-r}|\quad\text{for all } t>\nu.&\end{aligned} \tag{6.31}$$

Since $B(z)$ is stable, it follows from (6.31) and Lemma 4(i) that for some $A>0$ and $0<\rho<1$, we have with probability 1,

$$|y_t-\hat y_{t,M}-\eta_t|\leqq A\sup_{i\geqq\nu}\|\theta_{i-d,M}-\tilde\theta\|\sum_{i=\nu}^{t}\rho^{t-i}(|y_i|+|\varepsilon_i|)+A\rho^tU_\nu,\qquad t>\nu, \tag{6.32}$$

where U_ν is a nonnegative random variable.

Take any $0<\delta<1$. Since $\theta_t \to \tilde{\theta}$ almost surely, it follows from (6.32) that there exists a random variable τ_δ such that with probability 1,

$$|y_t - \hat{y}_{t,M} - \eta_t| \leqq \delta \max_{i \leqq t} (|y_i - \eta_i| + |\eta_i| + |\varepsilon_i|) + \delta \quad \text{for } t \geq \tau_\delta. \tag{6.33}$$

By choosing τ_δ large enough, we also have with probability 1 that

$$\hat{y}_{t,M} = y_t^* \quad \text{if } t \geqq \tau_\delta \quad \text{and } t-d \notin J, \tag{6.34}$$

in view of Lemma 9(ii) and (6.10), and that

$$|y_t - \eta_t| \leqq \delta K_t^{1/2} \quad \text{if } t \geqq \tau_\delta \quad \text{and } t-d \in J, \tag{6.35}$$

in view of (6.28). From (6.33)-(6.35), it follows that with probability 1,

$$|y_t - \eta_t| \leqq \delta K_t^{1/2} + |y_t^*| + \delta \max_{i \leqq t} |y_i - \eta_i| + \delta \max_{i \leqq t} (|\eta_i| + |\varepsilon_i| + 1), \quad t \geqq \tau_\delta,$$

and therefore

$$(1-\delta) \max_{\tau_\delta \leqq i \leqq t} |y_i - \eta_i| \leqq \max_{\tau_\delta \leqq i \leqq t} |y_i^*| + \delta K_t^{1/2} + O(1) \quad \text{as } t \to \infty. \tag{6.36}$$

Since δ can be arbitrarily small and since $\eta_t = O(1)$ almost surely, we obtain from (6.36) and (6.5) that

$$y_t^2 = o(K_t \vee (\log t)^\gamma) \quad \text{a.s.} \tag{6.37}$$

By (6.37) and Lemma 4(ii), $u_t^2 = o(K_{t+d} \vee (\log t)^\gamma)$ almost surely. Moreover, by (6.33) and (6.37), $\hat{y}_{t,M}^2 = o(K_t \vee (\log t)^\gamma)$ almost surely. □

LEMMA 11. (i) $\limsup_{n\to\infty} \sum_{i=1}^{n+d} (y_i - \hat{y}_{i,M} - \eta_i)^2 / \log n \leqq (2d-1)(p(d)+k+d-1+h) E\eta_d^2$ almost surely.

(ii) *Define the* $(p(d)+k+d-1+h) \times 1$ *vectors*

$$v_i = (0, \cdots, 0, 1, \overset{\longleftarrow h-i \longrightarrow}{0, \cdots, 0})', \; i = 1, \cdots, h, \quad v_{h+1} = \tilde{\theta}. \tag{6.38}$$

Then $v_1, \cdots, v_{h+1}$ *are linearly independent, and for* $i = 1, \cdots, h+1$,

$$\sum_{t=d+1}^{n} (v_i' \phi_{t-d,M})^2 = O\left(\log n + \sum_{t=1}^{n} (y_t^*)^2\right) \quad a.s.$$

Proof. (i) By (5.4b), $\hat{C}_n(q^{-1})\xi_n = \phi_{n,M}$, where $\hat{C}_t(z) = 1 + \hat{c}_{t,1} z + \cdots + \hat{c}_{t,h} z^h \to C(z)$ almost surely. Hence by Lemmas 4(ii) and 10,

$$\|\xi_n\|^2 = o(K_n \vee (\log n)^\gamma) \quad \text{a.s.} \tag{6.39}$$

Moreover, by (5.4d),

$$\lambda_{\min}(P_n^{-1}) \geqq \left(\sum_{i=\nu_1+1}^{n} i^{-1} \right) \sim \log n,$$

and therefore in view of (6.39) and (6.6),

$$\xi_n' P_n \xi_n \leqq \|\xi_n\|^2 / \lambda_{\min}(P_n^{-1}) = o(1) \quad \text{a.s.}$$

Hence we can apply Theorem 5 together with (6.40) and Lemma 8(i) to conclude that

$$\begin{aligned} \sum_{t=1}^{n+d} (y_t - \hat{y}_{t,M} - \eta_t)^2 \leqq{} & (2d-1)(E\eta_d^2 + o(1)) \log \det P_n^{-1} + o(\log n) \\ & + o(K_n^2 \vee K_n (\log n)^\gamma) \quad \text{a.s.} \end{aligned} \tag{6.40}$$

By (6.6) and (6.39), $\lambda_{\max}(P_{\nu_1}^{-1} + \sum_{t=\nu_1+1}^{n} \xi_t \xi_t') = o(n \log n)$ almost surely. Hence the desired conclusion follows.

(ii) Fix $i=1,\cdots,h$. From (6.38) and the definition of $\phi_{t,M}$ in (5.4b), it follows that $v_i'\phi_{t-d,M}=\hat{y}_{t-i,M}$. Combining this with (6.34) gives that with probability 1,

$$\sum_{t=d+1}^{n}(v_i'\phi_{t-d,M})^2\leqq\sum_{t=i+1}^{n}(y_{t-i}^*)^2+O(\#(J_{n+h})\sup_{t\leqq n-d}\|\phi_{t,M}\|^2)$$

$$\leqq\sum_{t=1}^{n}(y_t^*)^2+o(K_n^2\vee K_n(\log n)^\gamma),$$

by Lemmas 8(i) and 9. Moreover, by (6.34), with probability 1, for all large t,

$$v_{h+1}'\phi_{t-d,M}=\tilde{\theta}'\phi_{t-d,M}=(\tilde{\theta}-\theta_{t-d,M})'\phi_{t-d,M}+y_t^*\quad\text{if } t-d\notin J.$$

Since $(\tilde{\theta}-\theta_{t-d,M})'\phi_{t-d,M}=C(q^{-1})(y_t-\hat{y}_{t,M}-\eta_t)$ by Lemma 2, it then follows that

$$\sum_{t=d+1}^{n}(v_{h+1}'\phi_{t-d,M})^2=O\left(\sum_{t=1}^{n}(y_t-\hat{y}_{t,M}-\eta_t)^2+\sum_{t=1}^{n}(y_t^*)^2\right)+O(\#(J_n)\sup_{t\leqq n-d}\|\phi_{t,M}\|^2)$$

$$=O\left(\log n+\sum_{t=1}^{n}(y_t^*)^2\right)+o(K_n^2\vee K_n(\log n)^\gamma)\quad\text{a.s.}$$

by (i). Since $b_1\neq 0$, it is clear that $v_1,\cdots,v_{h+1}$ are linearly independent. □

Proof of Theorem 6. From (6.19), (6.6), and Lemmas 8(i) and 9(ii), it follows that

$$\sum_{i=1}^{n}(y_t-y_t^*-\eta_t)^2=\sum_{t\in L_n}(y_t-\hat{y}_{t,M}-\eta_t)^2+o(\log n)\quad\text{a.s.}\tag{6.41}$$

From (6.41) and Lemma 11(i), (6.11) follows.

Suppose furthermore that $\log(1+\sum_1^n(y_t^*)^2)=o(\log n)$ almost surely. Defining the linearly independent vectors $v_1,\cdots,v_{h+1}$ by (6.38), we obtain by Lemmas 4(ii) and 11(ii) that for $i=1,\cdots,h+1$,

$$\sum_{t\leqq n}(v_i'\xi_t)^2=O\left(\sum_{t\leqq n}(v_i'\phi_{t,M})^2\right)=O\left(\log n+\sum_{t=1}^{n+d}(y_t^*)^2\right)\quad\text{a.s.,}$$

noting that $\hat{C}_t(q^{-1})(v'\xi_t)=v'\phi_{t,M}$, where $\hat{C}_t(z)=1+\hat{c}_{t,1}z+\cdots+\hat{c}_{t,h}z^h\rightarrow C(z)$ almost surely. Since $P_n^{-1}=P_{\nu_1}^{-1}+\sum_{t=\nu_1+1}^n(\xi_t\xi_t'+t^{-1}I)$, it then follows that for $i=1,\cdots,h+1$,

$$\log(v_i'P_n^{-1}v_i)=\log\left\{v_i'P_{\nu_1}^{-1}v_i+\sum_{t=\nu_1+1}^{n}[(v_i'\xi_t)^2+t^{-1}\|v_i\|^2]\right\}=O(\log\log n)\quad\text{a.s.}\tag{6.42}$$

By (6.39), with probability 1, for all large n,

$$\log\operatorname{tr}(P_n^{-1})\leqq\log(n\log n)\sim\log n.\tag{6.43}$$

Combining (6.42) and (6.43), we can apply Lemma 5 to conclude that

$$\limsup_{n\to\infty}(\log\det P_n^{-1})/\log n\leqq(p(d)+k+d-1+h)-(h+1)=p(d)+k+d-2\quad\text{a.s.}$$

Therefore by (6.40),

$$\limsup_{n\to\infty}\sum_{t\leqq n}(y_t-\hat{y}_{t,M}-\eta_t)^2/\log n\leqq(2d-1)(p(d)+k+d-2)E\eta_d^2\quad\text{a.s.}$$

This and (6.41) give the desired conclusion (6.14). □

7. Extensions to multivariate systems and concluding remarks. Suppose that in (1.1), the y_n, ε_n, and u_{n-d} are $\nu\times 1$ vectors, and $A(q^{-1})$, $B(q^{-1})$, and $C(q^{-1})$ are matrix polynomials of the form (2.33) in the backward shift operator q^{-1}. As in (6.1), we will assume that $B(z)$ is a stable polynomial, i.e.,

(7.1) $$\det(B(z)) \neq 0 \quad \text{for all } |z| \leqq 1.$$

Moreover, analogous to (6.3)-(6.5), we will assume the following:

(7.2) The polynomials $z^p A(z^{-1})$, $z^{k-1}B(z^{-1})$ and $z^h C(z^{-1})$ are left coprime.

(7.3) $\{\varepsilon_n, \mathscr{G}_n, n \geqq 1\}$ is a martingale difference sequence such that $E(\varepsilon_n \varepsilon_n' \mid \mathscr{G}_{n-1}) = \Sigma$, a positive-definite nonrandom matrix, $\sup_n E(\|\varepsilon_n\|^\alpha \mid \mathscr{G}_{n-1}) < \infty$ almost surely for some $\alpha > 2$, and $\sup_n \|\varepsilon_n\| < \infty$ almost surely.

(7.4) y_n^* is an $\mathscr{F}_{n-d}$-measurable $\nu \times 1$ vector, $\sum_1^n \|y_i^*\|^2 = O(n)$ almost surely and $\|y_n^*\|^2 = o((\log n)^\gamma)$ for some $0 < \gamma < 1$.

In view of (7.1), $B_1 (= B(0))$ is nonsingular.

Let $\theta = (-A_1, \cdots, -A_p, B_1, \cdots, B_k, C_1, \cdots, C_h)'$, analogous to (1.6). In the unit-delay case $d = 1$, we can write (1.1) as a stochastic regression model (2.1) with $\psi_n = (y_n', \cdots, y_{n-p+1}', u_n', \cdots, u_{n-k+1}', \varepsilon_n', \cdots, \varepsilon_{n-h+1}')'$. The optimal controller assuming knowledge of θ and the initial condition x_0 determines the input u_n by the equation

(7.5) $$\theta' \psi_n = y_{n+1}^*.$$

In ignorance of θ and x_0, the so-called *explicit* (or *indirect*) approach of adaptive control is to replace θ in (7.5) by a recursive estimator θ_n and ψ_n by a pseudoregression vector ϕ_n that substitutes the unobservable ε_i in ψ_n either by the prediction error $y_i - \theta_{i-1}'\phi_{i-1}$ or by the residual $y_i - \theta_i'\phi_{i-1}$. In particular, Chen [27] used a modified least squares algorithm for θ_n, while Chen and Caines [28] used a stochastic gradient algorithm, and showed that the certainty-equivalence controller associated with either algorithm has the self-optimizing property (1.19) (with $R_n = \sum_2^N \|y_i - y_i^* - \varepsilon_i\|^2$) under certain assumptions.

In the unit-delay, single-input, single-output case (i.e., $d = 1 = \nu$), instead of the implicit approach as in § 6, we can use the explicit approach involving a parallel implementation of the stochastic gradient algorithms $\theta_{n,G}$, $\theta_{n,\tilde{G}}$ and the recursive maximum likelihood algorithm $\theta_{n,M}$ that are obvious modifications of (2.21), (3.7), and (5.4) for estimating the parameter vector (1.6) of the explicit model, instead of the parameter vector $\tilde{\theta}$ of the implicit model (2.17) considered in § 6. As shown in [16], there are exact analogues of Theorems 1 and 5 for the explicit case, and Theorems 2 and 3 can clearly be extended to the explicit case as well. The same argument also works for multivariable systems, for which we replace assumption (6.2) by

(7.6) $C(e^{it}) + C'(e^{-it}) - aI$ is positive definite for all $|t| \leqq \pi$ and some $a > 0$,

as has also been assumed by Chen and Caines [28]. Note that Lemma 3, which is a key tool in the proof of Theorem 2, has been stated for the multivariate positive real assumption (7.6). Moreover, Lemma 6, which is a key tool in the proof of Theorem 4, has also been stated under the multivariate coprime condition (7.2). Hence it is straightforward to generalize Theorem 4 to the multivariate case, and we can therefore obtain strongly consistent estimates of $(-A_1, \cdots, -A_p, B_1, \cdots, B_k)'$ by using the instrumental variables $Z_t = (y_{t-h-1}', \cdots, y_{t-h-k+1}', w_{t-1}', \cdots, w_{t-p-k}')'$, where the w_i satisfy the multivariate version of (3.5) and have a common positive-definite covariance matrix. A multivariate extension of the recursive implementation (4.24) of Wilson's [29] spectral factorization algorithm for multivariate time series can also be used to obtain strongly consistent estimates of $C_1, \cdots, C_h$, leading to the auxiliary estimator θ_{m_j} that satisfies (4.28).

Hence for unit-delay multivariate systems, under assumptions (7.1)-(7.4) and (7.6), we can modify the construction in § 6 (by using the explicit instead of the implicit approach) to obtain certainty-equivalence-type control rules that satisfy

$$\limsup_{n\to\infty} \sum_{i=2}^{n} \|y_i - y_i^* - \varepsilon_i\|^2/\log n \leqq \nu(p+k+h)\operatorname{tr}(\Sigma) \quad \text{a.s.} \tag{7.7}$$

Furthermore, if $\log(1+\sum_1^n \|y_i^*\|^2) = o(\log n)$ almost surely, then in analogy with (6.14), (7.7) can be strengthened as

$$\limsup_{n\to\infty} \sum_{i=2}^{n} \|y_i - y_i^* - \varepsilon_i\|^2/\log n \leqq \nu((p \vee h)+k-1)\operatorname{tr}(\Sigma) \quad \text{a.s.} \tag{7.8}$$

For the case of general delay, the construction in § 6 generalizes immediately to multivariable systems that satisfy either

$$A_j = a_j I \quad \text{for } j=1,\cdots,p, \tag{7.9}$$

so that $A(q^{-1})$ can be reduced to a scalar polynomial $1+a_1q^{-1}+\cdots+a_pq^{-p}$, or

$$C_j = c_j I \quad \text{for } j=1,\cdots,h, \tag{7.10}$$

so that $C(q^{-1})$ can be reduced to scalar polynomial $1+c_1q^{-1}+\cdots+c_hq^{-h}$. When (7.10) holds, $C(z)$ commutes with any $\nu\times\nu$ matrix polynomial and the minimum variance d-step ahead predictor $\tilde{y}_t \triangleq E(y_{t+d}\mid\mathcal{F}_t)$ can still be written in the form (2.14) with FB instead of BF, where $F(z)$, $G(z)$ are matrix polynomials uniquely determined by (2.12) (cf. [14, p. 135]). Hence a straightforward generalization of the construction in § 6 (using the implicit approach) leads to adaptive control rules that satisfy

$$\limsup_{n\to\infty} \sum_{i=1}^{n} \|y_i - y_i^* - \eta_i\|^2/\log n \leqq (2d-1)\{\nu(p(d)+k+d-1)+h\}E(\|\eta_d\|^2), \tag{7.11}$$

where $\eta_t = F(q^{-1})\varepsilon_t$, under assumptions (7.1)-(7.4) and (6.2), noting that $C(z)$ can be treated as a scalar polynomial in this case.

Case (7.9) has been studied by Goodwin, Ramadge, and Caines [8]. Let $\bar{c}(z) = \det(C(z))$ and $\tilde{C}(z)$ be the adjoint of $C(z)$. Let $F(z)$, $G(z)$ be the matrix polynomials uniquely determined by (2.12). Then (1.1) can be written in the following (d-step ahead) prediction form:

$$\bar{c}(q^{-1})[y_{n+d} - F(q^{-1})\varepsilon_{n+d}] = G(q^{-1})\tilde{C}(q^{-1})y_n + F(q^{-1})\tilde{C}(q^{-1})B(q^{-1})u_n \tag{7.12}$$

(cf. [8, pp. 852-853]). Hence, under assumptions (7.1)-(7.4) and

$$\operatorname{Re}\{\bar{c}(e^{it}) - (d-\tfrac{1}{2})a\} > 0 \quad \text{for all } |t| \leqq \pi \text{ and some } a>0, \tag{7.13}$$

a straightforward generalization of the construction in § 6 gives adaptive control rules whose regrets $\sum_{i=1}^n \|y_i - y_i^* - F(q^{-1})\varepsilon_i\|^2$ have logarithmic order. For $d=1$ (unit delay), assuming the positive real condition (7.13) together with some other conditions, Goodwin, Ramadge, and Caines [8] used the stochastic gradient algorithm to construct self-optimizing controllers in this case.

The logarithmic order for the regret of asymptotically efficient control rules also appears in the following classical problem of stochastic adaptive control. Consider k independent statistical populations specified respectively by univariate density functions $f(y;\theta_j)$ with finite means $\mu(\theta_j)$, $j=1,\cdots,k$, where the θ_j are unknown parameters belonging to some set Θ. How should we sample $y_1, y_2, \cdots$ sequentially from the k populations to maximize, in some sense, the expected value of the sum $S_n = y_1+\cdots+y_n$ as $n\to\infty$? This is the so-called "multi-armed bandit problem," whose name derives from an imagined slot machine with $k \geqq 2$ arms. When an arm is pulled, the player wins a random reward. For each arm j there is an unknown probability distribution Π_j of the reward. The player's problem is to choose a sequence of pulls on the k arms so as to maximize the long-run expected total reward (cf. [2], [30]).

Let $\theta = (\theta_1, \cdots, \theta_k)'$, $\mu^*(\theta) = \max_{j \leqq k} \mu(\theta_j)$. When θ is known, the optimal rule is to sample from the population whose mean attains $\mu^*(\theta)$. Without assuming θ to be known, an adaptive allocation rule u is a sequence of random variables $u_1, u_2, \cdots$, taking values in the set $\{1, \cdots, k\}$ and such that the event $\{u_n = j\}$ ("sample from Π_j at stage n") belongs to the σ-field $\mathscr{F}_{n-1}$ generated by the previous observations $u_1, y_1, \cdots, u_{n-1}, y_{n-1}$. Robbins [30] formulated a notion of "asymptotic optimality" for an adaptive allocation rule u as attaining

$$\lim_{n \to \infty} n^{-1} E_\theta S_n = \mu^*(\theta) \quad \text{for all } \theta \in \Theta. \tag{7.14}$$

This is analogous to the "self-optimizing" criterion (1.19′) for adaptive controllers in the ARMAX system (1.1). In the case $k = 2$, Robbins introduced the following class of certainty-equivalence rules with forced learning to achieve (7.14). Sample from Π_1 during stages $n_1 < n_2 < \cdots$ and from Π_2 during stages $n_1^* < n_2^* < \cdots$, where $\{n_i\}$ and $\{n_i^*\}$ are disjoint sequences of positive from Π_2 during stages $n_1^* < n_2^* < \cdots$, where $\{n_i\}$ and $\{n_i^*\}$ are disjoint sequences of positive integers such that $n_i / i \to \infty$ and $n_i^* / i \to \infty$. At stage $n \notin \{n_1, n_2, \cdots, n_1^*, n_2^*, \cdots\}$, sample from the population with the larger sample mean.

For an adaptive allocation rule, $E_\theta S_n = \sum_{j=1}^k \mu(\theta_j) E_\theta T_n(j)$, where $T_n(j)$ denotes the total number of observations from Π_j up to stage n (cf. [31]). Thus, maximizing $E_\theta S_n$ is equivalent to minimizing $E_\theta R_n$, where we define the regret R_n by

$$R_n = \sum_{j: \mu(\theta_j) < \mu^*(\theta)} (\mu^*(\theta) - \mu(\theta_j)) T_n(j), \tag{7.15}$$

which is a weighted sum of the sample sizes from the "inferior" populations. Note the analogy between (7.15) and (1.16). Lai and Robbins [31] recently showed that for any adaptive allocation rule satisfying $E_\theta R_n = o(n^a)$ for every $a > 0$ and every θ, we also have

$$\liminf_{n \to \infty} E_\theta R_n / \log n \geqq \sum_{j: \mu(\theta_j) < \mu^*(\theta)} (\mu^*(\theta) - \mu(\theta_j)) / I(\theta, \theta^*), \tag{7.16}$$

where $\mu(\theta^*) = \max_{j \leqq k} \mu(\theta_j)$ and $I(\theta, \lambda)$ denote the Kullback-Leibler information number. Moreover, adaptive allocation rules that attain the asymptotic lower bound in (7.16) at every θ are constructed in [31] for various parametric families of distributions. Such rules involve certain upper confidence bounds for $\mu(\theta_j)$ and are typically of the form: Sample at each stage from the population with the largest upper confidence bound. These upper confidence bounds are constructed by using generalized likelihood ratios, analogous to the maximum likelihood approach of § 5 that has been used in the construction of asymptotically efficient adaptive controllers for the ARMAX model in § 6.

As pointed out in [32], a common difficulty in the multi-armed bandit problem, the adaptive control problem in ARMAX systems studied herein, and the multiperiod control problem of choosing the design levels u_i in the regression model (1.20) to minimize $E(\sum_1^N y_i^2)$ is the dilemma between the object of efficient control and the need for information in estimating the system parameters. Although a Bayesian formulation (involving a suitable prior distribution on the unknown parameters) and the associated Bayes solution should, in principle, be able to resolve this dilemma, the Bayesian optimal control problem is prohibitively difficult to solve either numerically or analytically. A useful heuristic principle described in [32] to avoid these difficulties is to consider the fictitious situation that assumes knowledge of some crucial parameter(s) so that there is negligible conflict between estimation and control, leading to a more tractable Bayes problem. Analysis of this fictitious situation yields the asymptotic lower bound (7.16) for the expected regret in the multi-armed bandit problem, the logarithmic order (1.23) of the regret of the optimal rule in the multiperiod control problem for the regression model (1.20) with known β, and the asymptotic lower bound (1.14) for the Bayes regret in the regulation problem for the ARX model that assumes b_1 to be known (cf. [4], [5], [32]). The ensuing task, therefore, is to develop recursive control schemes whose regrets do not exceed asymptotically those in (7.16), (1.23), or (1.14), even when *all* parameters are unknown.

For the adaptive control problem of ARMAX systems, we have shown in § 6 that a modified version of the simple certainty-equivalence idea is powerful enough to accomplish this task. Since the goal is to emulate the performance of the Bayes rule, which is too complex even for off-line implementation, it is intuitively clear that a statistically efficient estimator should be used for the certainty-equivalent rule. This explains the use of the monitored recursive maximum likelihood algorithm introduced in § 5. Except in the white-noise case $C(z)=1$, the estimating equations defining the off-line maximum likelihood estimators (assuming independently and identically distributed Gaussian disturbances ε_i) are highly nonlinear. However, when monitored by an auxiliary *consistent* estimator, the recursive maximum likelihood estimator $\theta_{n,M}$ (which corresponds at every stage to a one-step Gauss–Newton iteration to solve the off-line estimating equation) is asymptotically as efficient as its off-line counterpart (cf. [33]). The need for an auxiliary consistent estimator (which we construct by the method of moments applied to blocks of well-excited input-output data) accounts for the modification (3.8) of the certainty-equivalence rule based on $\theta_{n,M}$. During the initial stages when the auxiliary consistent estimator may not be sufficiently reliable, it seems more prudent to use the stochastic gradient algorithm $\theta_{n,G}$ (which is known to have a stabilizing effect) instead of the (possibly not well monitored) $\theta_{n,M}$ in the certainty-equivalence rule. This is the rationale behind the additional modification (6.9) of the obvious certainty-equivalence rule based on $\theta_{n,M}$. Since the algorithms $\theta_{n,M}$, $\theta_{n,G}$, and $\theta_{n,\tilde{G}}$ used in the certainty-equivalence inputs of Theorem 6 are all recursive, running them in parallel does not cause difficulties for on-line implementation.

As shown in [33], the classical method of moments, which involves only a fixed number of auto/cross-covariances that can be easily updated, offers a reliable and versatile way to construct consistent recursive estimators in ARMAX systems but has not been fully explored in the recursive identification literature. Although the method of moments and its application only to certain segments of all available data are inefficient statistical procedures, they provide valuable initial parameter estimates around which we can linearize the complicated nonlinear estimating equations that define the statistically efficient but computationally intensive off-line maximum likelihood estimators. Such linearization leads to the monitored recursive maximum likelihood estimator of § 5, which is asymptotically as efficient as its off-line counterpart. Note in this connection that although the AML algorithm (2.6) has been called "approximate maximum likelihood," its statistical properties are actually unrelated to those of the off-line maximum likelihood estimator associated with Gaussian ε_i. Although it has some resemblance to the iterative EM algorithm that is sometimes used to compute the off-line maximum likelihood estimator, the AML algorithm is nothing more than a formal extension of the recursive least squares algorithm (2.2) and can run into serious difficulties when $C(z)$ differs so much from 1 (the white-noise case) that the positive real assumption (2.8) is violated (cf. [33]).

REFERENCES

[1] K. J. ÅSTRÖM, *Theory and applications of adaptive control—A survey*, Automatica—J. IFAC, 19 (1983), pp. 471-486.

[2] P. R. KUMAR, *A survey of some results in stochastic adaptive control*, SIAM J. Control Optim., 23 (1985), pp. 329-380.

[3] A. A. FELDBAUM, *The theory of dual control* I-IV, Automat. Remote Control, 21 (1961), pp. 874-883 (Part I) and pp. 1033-1039 (Part II), 22 (1962), pp. 1-12 (Part III), and pp. 109-121 (Part IV).

[4] T. L. LAI, *Asymptotically efficient adaptive control in stochastic regression models*, Adv. in Appl. Math., 7 (1986), pp. 23-45.

[5] T. L. LAI AND H. ROBBINS, *Adaptive design and stochastic approximation*, Ann. Statist., 7 (1979), pp. 1196-1221.

[6] R. E. KALMAN, *Design of a self-optimizing control system*, Trans. ASME, 80 (1958), pp. 468-478.

[7] K. J. ÅSTRÖM AND B. WITTENMARK, *On self-tuning regulators*, Automatica—J. IFAC, 9 (1973), pp. 195-199.

[8] G. C. GOODWIN, P. J. RAMADGE, AND P. E. CAINES, *Discrete time stochastic adaptive control*, SIAM J. Control Optim., 19(1981), pp. 829-853.

[9] P. E. CAINES AND S. LAFORTUNE, *Adaptive control with recursive identification for stochastic linear systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 312-321.

[10] K. S. SIN AND G. C. GOODWIN, *Stochastic adaptive control using a modified least squares algorithm*, Automatica—J. IFAC, 18 (1982), pp. 315-321.

[11] T. L. LAI AND C. Z. WEI, *Asymptotically efficient self-tuning reulators*, SIAM J. Control Optim., 25 (1987), pp. 466-481.

[12] ———, *Extended least squares and their applications to adaptive control and prediction in linear systems*, IEEE Trans. Automat. Control, 31 (1986), pp. 898-906.

[13] G. C. GOODWIN AND K. S. SIN, *Adaptive Filtering, Prediction and Control*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[14] P. E. CAINES, *Linear Stochastic Systems*, John Wiley, New York, 1988.

[15] G. C. GOODWIN, K. S. SIN, AND K. K. SALUJA, *Stochastic adaptive control and prediction—the general delay-colored noise case*, IEEE Trans. Automat. Control, 25 (1980), pp. 946-949.

[16] T. L. LAI AND Z. YING, *Recursive indentification and adaptive prediction in linear stochastic systems*, SIAM J. Control Optim., 29 (1991), pp. 1061-1090.

[17] H. J. KUSHNER, *Stochastic Stability and Control*, Academic Press, New York, 1967.

[18] R. Z. HAS'MINSKII, *Stochastic Stability of Differential Equations*, Sijthoff and Noordhoff, the Netherlands, 1980.

[19] T. L. LAI AND C. Z. WEI. *On the concept of excitation in least squares identification and adaptive control*, Stochastics, 16 (1986), pp. 227-254.

[20] V. SOLO, *The convergence of* AML, IEEE Trans. Automat. Control, 24 (1979), pp. 958-962.

[21] S. P. MEYN AND P. E. CAINES, *The zero divisor problem of multivariable stochastic adaptive control*, System Control Lett., 6 (1985), pp. 235-238.

[22] G. WILSON, *Factorization of the covariance generating function of a pure moving average process*, SIAM J. Numer. Anal., 6 (1969), pp. 1-7.

[23] V. SOLO, *Adaptive spectral factorization*, IEEE Trans. Automat. Control, 34 (1989), pp. 1047-1051.

[24] T. L. LAI AND Z. YING, *Recursive solutions of nonlinear stochastic equations with applications to spectral factorization*, IEEE Trans. Automat. Control, to appear.

[25] L. LJUNG AND T. SÖDERSTRÖM, *Theory and Practice of Recursive Identification*, MIT Press, Cambridge, MA, 1983.

[26] M. S. BAZARAA AND C. M. SHETTY, *Nonlinear Programming—Theory and Algorithms*, John Wiley, New York, 1979.

[27] H. F. CHEN, *Recursive system identification and adaptive control by use of the modified least squares algorithm*, SIAM J. Control Optim., 22 (1984), pp. 758-776.

[28] H. F. CHEN AND P. E. CAINES, *The strong consistency of the stochastic gradient algorithm of adaptive control*, in Proc. 23rd IEEE Conference on Decision and Control, Las Vegas, NV, 1984, pp. 802-805.

[29] G. T. Wilson, *A convergence theorem for spectral factorization*, J. Multivariate Anal., 8 (1978), pp. 222-232.

[30] H. Robbins, *Some aspects of the sequential design of experiments*, Bull. Amer. Math. Soc., 55 (1952), pp. 527-535.

[31] T. L. Lai and H. Robbins, *Asymptotically efficient adaptive allocation rules*, Adv. in Appl. Math., 6 (1985), pp. 4-22.

[32] T. L. Lai, *Some thoughts on stochastic adaptive control*, in Proc. 23rd IEEE Conference on Decision and Control, Las Vegas, NV, 1984, pp. 51-56.

[33] ———, *Recursive estimation in* ARMAX *models*, in New Directions in Time Series, E. Parzen et al., eds., Springer-Verlag, New York, to appear.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 5, pp. 1128-1149, September 1991

006

SENSITIVITY ANALYSIS OF OPTIMAL CONTROL PROBLEMS FOR WAVE EQUATIONS*

IRENA LASIECKA† AND JAN SOKOŁOWSKI‡

Abstract. Differential stability of the optimal solutions to an optimal control problem for hyperbolic partial differential equation is shown. The related results on the sensitivity analysis for hyperbolic equation with respect to perturbations of the coefficients of elliptic operator are provided.

Key words. wave equation, metric projection, sensitivity analysis, control constraints

AMS(MOS) subject classifications. 49B22, 49A29, 49A22, 93B30

1. Introduction. This paper is devoted to the sensitivity analysis of the boundary control constrained optimal control problems for the wave equation. We use the method proposed in [S6] combined with the recent results on the regularity of solutions to the wave equation [LLT], [LT1]. We consider for simplicity a quadratic problem; however, the same argument can be used in the convex case.

Since, in general, the model of the distributed parameter system described by the wave equation may not be known exactly, our results can be used to evaluate the increments of an optimal control corresponding to the given increments of the data, e.g., variable coefficients of the partial differential equation, or to the perturbation of the domain of integration.

For a specific convex optimal control problem any optimal control satisfies the necessary and sufficient optimality conditions in the form of an optimality system [L2]. By inspection of the optimality system it follows that in the case of control constraints, an optimal control is given by the unique fixed point of a metric projection onto the set of admissible constraints. This particular form of the optimal control allows us to obtain the right derivative of an optimal control with respect to the parameter. To this end the notion of Hadamard directional differentiability of the metric projection onto the polyhedric convex set in Hilbert space [H1], [M], is used. It seems that the same method of sensitivity analysis can be used for some nonconvex optimization problems such that the related results on the Hölder continuity of an optimal solution with respect to the parameter are known, such problems are the subject of a paper in preparation. There are several papers and monographs concerning the sensitivity analysis of finite-dimensional problems, in particular, in mathematical programming. On the other hand, the infinite-dimensional case requires the knowledge of the new results on the differential stability of metric projection onto the specific convex, closed subsets of Sobolev spaces. Such results are known in general only for a class of polyhedric sets [H1], [M], [RS2] defined by specific local constraints.

We provide here the results on sensitivity analysis for the wave equation which, we believe, is interesting on its own. We restrict ourselves in the present paper to a model problem. However, the method proposed here is general and can be used for a broad class of control problems including a class of state constrained optimal control problems, such as those considered in [RS1] in the elliptic case.

* Received by the editors June 20, 1988; accepted for publication (in revised form) November 30, 1990.

† Applied Mathematics Department, University of Virginia, Charlottesville, Virginia 22903. This research was partially supported by the National Science Foundation grant DMS-8902811 and by the Air Force Office of Scientific Research under grant AFOSR-87-0321.

‡ Systems Research Institute, Polish Academy of Sciences, ul. Newelska 6, 01-447 Warszawa, Poland. The author was sponsored by the Research Program of Polish Academy of Sciences, R.P.I.02.25.

Consider the following optimal control problem depending on parameter $\varepsilon \in [0, \delta)$.

PROBLEM (P_ε). Find an element $u_\varepsilon \in K$ that minimizes the cost functional

$$I_\varepsilon(u) = \tfrac{1}{2}(Ry(T), y(T))_\Omega + \tfrac{1}{2}|u|_\Sigma^2 \tag{1.1}$$

over the convex set $K \subset L^2(\Sigma)$ of the form

$$K = \{u \in L^2(\Sigma) \,|\, 0 \leqq u(x, t) \leqq M \text{ for a.e. } (x, t) \in \Sigma\}. \tag{1.2}$$

Here by y we denote the solution of the state equation

$$y_{tt} = A_\varepsilon(x, \partial)y \quad \text{in } \Omega \times (0, T) \tag{1.3}$$

with the Neumann boundary condition

$$\frac{\partial y}{\partial n} = u \quad \text{on } \Sigma \tag{1.4}$$

subject to initial conditions

$$y(0) = y_0, \quad y_t(0) = y_1 \quad \text{in } \Omega, \tag{1.5}$$

where $u \in L^2(\Sigma)$ is control, and $y_0 \in L^2(\Omega)$ and $y_1 \in H^1(\Omega)$ are given elements. Here we have used the following notation:

$$A_\varepsilon(x, \partial)y = \operatorname{div}(a_\varepsilon \nabla y) \quad \forall y \in H^2(\Omega),$$

$$\varepsilon \in [0, \delta), \quad \delta > 0 \quad \text{is parameter},$$

$$\Omega \subset R^n, \quad Q = \Omega \times (0, T), \quad T > 0,$$

$$\Gamma = \partial\Omega, \qquad \Sigma = \partial\Omega \times (0, T),$$

$$\nabla y = \left(\frac{\partial y}{\partial x_1}, \cdots, \frac{\partial y}{\partial x_n}\right),$$

$$\mathbf{z} = (z_1, \cdots, z_n),$$

$$|\phi|_\Omega^2 = \int_\Omega [\phi(x)]^2\, dx,$$

$$|\phi|_\Gamma^2 = \int_\Gamma [\phi(x)]^2\, d\sigma,$$

$$|\phi|_\Sigma^2 = \int_0^T |\phi(x)|_\Gamma^2\, dt.$$

$H^2(\Omega)$ for $s \geqq 0$ denotes the usual Sobolev space of order s. $H^{-s}(\Omega) \equiv (H^s(\Omega))'$ where the duality is understood with respect to the pivot space $L^2(\Omega)$.

The following assumptions are imposed on the data of the problem: $\Omega \subset R^n$ is a bounded, open domain with the smooth boundary Γ. The coefficients $a_\varepsilon(x)$ are given by (for each $\varepsilon > 0$)

$$a_\varepsilon(x) = a_0(x) + \varepsilon a_1(x) + o(\varepsilon)(x) \quad \text{in } C^1(\Omega), \tag{1.6}$$

where $a_0(x) \geqq c > 0$ in Ω

$$R: L^2(\Omega) \to L^2(\Omega), \qquad R = R^* > 0$$

and either

(1.7) (i) $R \in \mathscr{L}(H^s(\Omega))$, $0 < s < \frac{1}{2}$ and $y_0 \in H^{1/2}(\Omega)$; $y_1 \in (H^{1/2}(\Omega))' = H^{-1/2}(\Omega)$, or

(1.7) (ii) $R \in \mathscr{L}(H^s(\Omega); H^{s-1}(\Omega))$, $1 \leqq s \leqq \frac{3}{2}$ and $y_0 \in H^{3/2}(\Omega)$, $y_1 \in H^{1/2}(\Omega)$.

It is standard to show that the optimal solution u_ε to the problem (P_ε) exists, is unique, and

$$|u_\varepsilon|_\Sigma \leqq C \quad \text{uniformly in } \varepsilon > 0.$$

The same holds true for unbounded set (1.2) of admissible controls, i.e., for $M = +\infty$.

Our main goal in this paper is to characterize the right-derivative q of the unique solution $u_\varepsilon \in K$ of Problem (P_ε) with respect to the parameter ε at $\varepsilon = 0$.

Several results on differential stability of solutions to control problems for elliptic and parabolic partial differential equations are given in [S1]-[S7] and [MS]. Problems with the state constraints are considered in [RS1]. The related results on sensitivity analysis of the variational inequalities can be found in [M], [H-S-Z], and [FP].

In the hyperbolic case, the main difficulty is related to the intrinsic "low" regularity of the solutions with the boundary inputs in $L^2(\Sigma)$. In fact, standard regularity theory for the wave equation with boundary nonhomogenous data (see [LM]) is inadequate to obtain the results on differential stability for Problem (P_ε). To cope with the problem, we shall use the recently obtained (in [LT1]) "sharp" regularity results for the Neumann problem, which will play a crucial role in our analysis.

Our main result is formulated in the theorem below.

THEOREM 1. *For each $\varepsilon > 0$, ε small enough*

$$u_\varepsilon = u_0 + \varepsilon q + o(\varepsilon) \quad \text{in } L^2(\Sigma), \tag{1.8}$$

where $|o(\varepsilon)|_\Sigma/\varepsilon \to 0$ with $\varepsilon \downarrow 0$.

The element $q \in L^2(\Sigma)$ is given by a unique solution of the following optimal control problem.

PROBLEM (Q). Find an element $q \in S \subset L^2(\Sigma)$ that minimizes the cost functional

$$J(u) = \tfrac{1}{2}(Rz(T), z(T))_\Omega + \tfrac{1}{2}|u|_\Sigma^2 \tag{1.9}$$

over cone

$$\begin{aligned} S = \{u \in L^2(\Sigma) \mid\, & u(x) \geqq 0 \text{ a.e. on } \{u_0(x) = 0\}, \\ & u(x) \leqq 0 \text{ a.e. on } \{u_0(x) = M\}, \\ & \textstyle\int_\Sigma (u_0(x,t) + a_0(x)p_0(x,t))u(x,t)\, d\Sigma = 0\}. \end{aligned} \tag{1.10}$$

Here we denote by $p_0 \in L^2(\Sigma)$ the trace on Σ of the adjoint state associated to the problem (P_0) (i.e., (P_ε) with $\varepsilon = 0$), i.e., $p \equiv p_0$ satisfies

$$p_{tt} = A(x, \partial)p, \qquad \frac{\partial}{\partial n} p = 0,$$

$$p(T) = 0, \qquad p_t(T) = Ry^0(T).$$

The state z for Problem (Q) is given for a given control $u \in L^2(\Sigma)$ by the unique solution of the following system:

$$z_{tt} = \operatorname{div}(a_0 \nabla z + a_1 \nabla y^0) \quad \text{in } Q, \tag{1.11}$$

$$\frac{\partial z}{\partial n} = u + \frac{a_1}{a_0} u_0 \quad \text{on } \Sigma, \tag{1.12}$$

$$z(0) = 0, \quad z_t(0) = 0 \quad \text{in } \Omega, \tag{1.13}$$

where y^0 is the solution of system (1.3)-(1.5) for $\varepsilon = 0$ and for $u = u_0$.

By standard argument it follows that an optimal control u_ε is uniquely determined by the optimality system [L2]. Let us recall that to derive the optimality system for Problem (P_ε) we define the Lagrangian

$$L_\varepsilon(u, y, p) \equiv -(y_t, p_t)_Q + (y_t(T), p(T))_\Omega - (y_1, p(0))_\Omega$$
$$+ (a_\varepsilon \nabla y, \nabla p)_Q - (u, a_\varepsilon p)_\Sigma + \tfrac{1}{2}|u|_\Sigma^2 + \tfrac{1}{2}(Ry(T), y(T))_\Omega, \qquad y(0) = y_0.$$

The first-order optimality conditions

$$\langle \mathscr{D}_y L_\varepsilon(u_\varepsilon, y_\varepsilon, p_\varepsilon); \delta y\rangle_Q = 0 \quad \forall \delta y,$$
$$\langle \mathscr{D}_p L_\varepsilon(u_\varepsilon, y_\varepsilon, p_\varepsilon); \delta p\rangle_Q = 0 \quad \forall \delta p,$$
$$u_\varepsilon \in K: \quad (\mathscr{D}_u L_\varepsilon(u_\varepsilon, y_\varepsilon, p_\varepsilon), v - u_\varepsilon)_\Sigma \geqq 0 \quad \forall v \in K,$$

where δy, $\delta p \in H^1(0, T; L^2(\Omega)) \cap L^2(0, T; H^1(\Omega))$, $\delta y(0) = 0$, $\delta y_t(0) = 0$ lead to the following optimality system.

Find $(u_\varepsilon, y^\varepsilon, p^\varepsilon)$ such that the following system is satisfied:

State equation:

$$y^\varepsilon_{tt} = A_\varepsilon(x, \partial) y^\varepsilon \quad \text{in } Q, \tag{1.14}$$

$$\frac{\partial y^\varepsilon}{\partial n} = u_\varepsilon \quad \text{on } \Sigma, \tag{1.15}$$

$$y^\varepsilon(0) = y_0, \quad y^\varepsilon_t(0) = y_1 \quad \text{in } \Omega. \tag{1.16}$$

Adjoint state equation:

$$p^\varepsilon_{tt} = A_\varepsilon(x, \partial) p^\varepsilon \quad \text{in } Q, \tag{1.17}$$

$$\frac{\partial p^\varepsilon}{\partial n} = 0 \quad \text{on } \Sigma, \tag{1.18}$$

$$p^\varepsilon(T) = 0, \quad p^\varepsilon_t(T) = R y_\varepsilon(T) \quad \text{in } \Omega. \tag{1.19}$$

Optimality conditions:

$$u_\varepsilon \in K: \quad \int_\Sigma (u_\varepsilon + a_\varepsilon p_\varepsilon)(v - u_\varepsilon)\, d\Sigma \geqq 0 \quad \forall v \in K. \tag{1.20}$$

We see that an optimal control u_ε takes the form

$$u_\varepsilon = \Pi(-\gamma a_\varepsilon p_\varepsilon) = \Pi(-\{a_\varepsilon p_\varepsilon\}_{|\Sigma});$$

here $\gamma \in \mathscr{L}(H^s(\Omega); H^{s-1/2}(\Gamma))$, $s > \frac{1}{2}$ is the trace operator, Π denotes the metric projection in $L^2(\Sigma)$ onto the convex set (1.2); the explicit form of Π is given by (2.17).

The outline of the paper is as follows. In § 2 we recall the basic results for the wave equation that we need in the sequel. We also present a result on the directional differentiability of the metric projection onto the set of admissible controls.

In § 3 we formulate the results on the sensitivity analysis for the wave equation and prove the main result of the paper. In §§ 4 and 5 we provide the proofs of the results on the sensitivity analysis of the solutions of the state equation used for the Theorem 1.

2. Preliminaries.

2.1. Wave equation; regularity of the solutions. We will consider the optimal control problems for hyperbolic problems. First we recall the basic facts on hyperbolic initial boundary value problems that we use in the sequel.

Let us consider the wave equation

$$y_{tt} = \operatorname{div}(a_0 \nabla y) \quad \text{in } Q, \tag{2.1}$$

$$\frac{\partial y}{\partial n} = u \quad \text{on } \Sigma, \tag{2.2}$$

$$y(0) = y_0, \quad y_t(0) = y_1 \quad \text{in } \Omega, \tag{2.3}$$

where $a_0(\cdot) \in C^1(\bar{\Omega})$ is a given coefficient and $u(\cdot) \in L^2(\Sigma)$. We assume that $a_0(x) \geqq c > 0$, $x \in \bar{\Omega}$. The initial conditions $y_0 \in L^2(\Omega)$, $y_1 \in H^{-1}(\Omega)$. The solution $y(\cdot)$ of system (2.1)-(2.3) can be represented (see [LT2]) in the form

$$y(t) = (Lu)(t) + C(t)y_0 + S(t)y_1, \tag{2.4}$$

where the linear mapping

$$L: L^2(\Sigma) \rightarrow L^2(Q)$$

takes the form

$$(Lu)(t) = A \int_0^t S(t-\tau)(Nu)(\tau)\, d\tau - \int_0^t S(t-\tau)(Nu)(\tau)\, d\tau. \tag{2.5}$$

Here $A: L^2(\Omega \rightarrow L^2(\Omega))$ denotes the generator of the analytic semigroup associated with $A(x, \partial)$ and zero Neumann boundary conditions with

$$D(A) \equiv \{y \in H^2(\Omega);\ \partial y/\partial n|_\Gamma = 0\},$$

$C(t)$ and $S(t)$ are cosine and sine operators associated with A, and the Neumann operator is defined as follows:

$$N: L_2(\Gamma) \rightarrow L_2(\Omega),$$

$$-\operatorname{div}(a_0 \nabla(Nu)) + Nu = 0 \quad \text{in } \Omega, \tag{2.6}$$

$$\frac{\partial(Nu)}{\partial n} = u \quad \text{on } \Gamma. \tag{2.7}$$

By applying the regularity results of [LM] we obtain

$$L \in \mathscr{L}(L^2(\Sigma);\ C[0, T;\ H^{1/2}(\Omega)]). \tag{2.8}$$

Hence the adjoint (adjoint in the sense of L^2-topology) operator

$$L^* \in \mathscr{L}(L^1[0, T;\ (H^{1/2}(\Omega))'];\ L^2(\Sigma)), \tag{2.9}$$

where L^* is defined as follows:

$$(L^*f)(t) = \int_t^T N^*A^*S^*(\tau - t)f(\tau)\, d\tau = a_0\left(\int_t^T S^*(\tau - t)f(\tau)\, d\tau\right)\bigg|_\Gamma. \tag{2.10}$$

To obtain (2.10), we use

$$N^*A^*z = a_0 z|_\Gamma \quad \forall z \in D(A^*), \tag{2.11}$$

which can be obtained using Green's formula.

We will also need some properties of the linear mapping

$$L_T^*: L^2(\Sigma) \rightarrow (Lu)(T) \in D(A^{1/4}) = H^{1/2}(\Omega) \tag{2.12}$$

and of the adjoint mapping

$$L_T^*: D(A^{1/4})' \rightarrow L^2(\Sigma), \tag{2.13}$$

where

$$(L_T^*\psi)(t) = a_0(S^*(T-t)\psi)|_\Gamma \quad \forall \psi \in L^2(\Omega)$$

and X' stands for the dual space to X with respect to $L_2(\Omega)$ topology. To obtain (2.12), we have used the following identification:

$$\text{(2.14)} \qquad D(A^Q) = H^{2Q}(\Omega), \qquad 0 \leqq Q < \tfrac{3}{4}.$$

In the case where the coefficient a_0 in (2.1) is replaced by $a_\varepsilon \in C^1(\bar{\Omega})$ all the results remain valid with the estimates independent on $0 \leqq \varepsilon \leqq \delta$ in (2.8) and (2.12).

The partial differential equation interpretation of (2.10) is as follows:

$$(L^*f)(t) = a_0 w(t)_{|\Gamma},$$

where $w(t)$ satisfies

$$w_{tt} = A(x, \partial)w + f(t) \quad \text{in } Q,$$

$$\frac{\partial w}{\partial n} = 0 \quad \text{on } \Sigma,$$

$$w(T) = w_t(T) = 0 \quad \text{in } \Omega.$$

Similarly for (2.13),

$$(L_T^* y)(t) = a_0 z(t)_{|\Gamma}$$

where $z(t)$ satisfies

$$z_{tt} = A(x, \partial)z \quad \text{in } Q,$$

$$\frac{\partial z}{\partial n} = 0 \quad \text{on } \Sigma,$$

$$z(T) = 0, \quad z_t(T) = y \quad \text{in } \Omega.$$

In the sequel we will need stronger regularity results than those given in (2.8) and (2.12). In fact we will need "sharp" regularity results, which are collected in the lemma below.

LEMMA 1 [LT1]. *Let ρ be arbitrarily small.*

(i) $L_T \in \mathscr{L}(L^2(\Sigma); D(A^{1/4+\alpha})) \cap \mathscr{L}(H^{1-2\alpha-2\rho}(\Sigma); D(A^{3/4-\rho}))$;

(ii) $L_T^* \in \mathscr{L}((D(A^{1/4+\alpha}))'; L^2(\Sigma)) \cap \mathscr{L}(D(A^{1/4-\rho}); H^{1+2\alpha-2\rho}(\Sigma))$;

(iii) $L \in \mathscr{L}(L^2(\Sigma); C[0, T; D(A^{1/4+\alpha})]) \cap \mathscr{L}(H^{1-2\alpha-2\rho}(\Sigma); C[0, T; D(A^{3/4-\rho})])$;

(iv) $L \in \mathscr{L}(H^s(\Sigma); H^{s+1/2+2\alpha}(Q))$ *provided that for* $s + \frac{1}{2} + 2\alpha \geqq \frac{3}{2}$, $0 \leqq s \leqq 1$, *the compatibility condition* $u(0) = 0$ *on* Γ *is satisfied.*

Here $1/20 \leqq \alpha \leqq \alpha_0(\Omega)$ with $\alpha_0(\Omega)$ depending, in general, on the geometry of the domain Ω.

Remark 1. In [LT1] it was shown that for smooth domains Ω we have $\alpha_0(\Omega) \geqq 1/20$. However, in the special cases of the domain Ω, the value of $\alpha_0(\Omega)$ may be bigger. For example, if Ω is a sphere (respectively, parallelepiped) then $\alpha_0(\Omega) = 1/6$ (respectively, $\frac{1}{4}$). In any case, the results of Lemma 1 state that the regularity of the solutions to the wave equation with Neumann boundary data is "at least" "1/10" derivative higher than the regularity predicted by the standard theory. In our applications, however, the exact value of α_0 is not important. It is crucial, however, that $\alpha_0 > 0$.

2.2. Projection in $L^2(\Sigma)$. Let $K \subset L^2(\Sigma)$ be a set of the form

$$K = \{v \in L^2(\Sigma) \mid 0 \leqq v(x, t) \leqq M, (x, t) \in \Sigma\}. \tag{2.15}$$

Let us denote by $\Pi: L^2(\Sigma) \to L^2(\Sigma)$ the metric projection in $L^2(\Sigma)$ onto the set of admissible controls K, i.e.,

$$\forall f \in L^2(\Sigma): \quad |\Pi f - f|_\Sigma = \min_{v \in K} |v - f|_\Sigma. \tag{2.16}$$

It can be shown that

$$(\Pi f)(x, t) = \min \{M, \max \{f(x, t), 0\}\}; \tag{2.17}$$

therefore,

$$(\Pi f)(x, t) = F(f(x, t)), \tag{2.18}$$

where

$$F(r) = \begin{cases} M, & r > M, \\ r, & 0 \leqq r \leqq M, \\ 0, & r < 0. \end{cases} \tag{2.19}$$

Hence it follows that

$$\Pi: H^{s,s}(\Sigma) \to H^{s,s}(\Sigma) \tag{2.20}$$

is continuous for all $s \in [0, 1]$.

We also have the following result on the directional differentiability of the mapping:

$$\Pi: L^2(\Sigma) \to L^2(\Sigma).$$

LEMMA 2. *For $\tau > 0$, τ small enough*

$$\forall h \in L^2(\Sigma): \quad \Pi(f + \tau h) = \Pi(f) + \tau \Pi'_f h + o(\tau), \tag{2.21}$$

where $\|o(\tau)\|_{L^2(\Sigma)}/\tau \to 0$ with $\tau \downarrow 0$ and $\Pi'_f: L^2(\Sigma) \to L^2(\Sigma)$ is the metric projection in $L^2(\Sigma)$ onto cone

$$\begin{aligned} S_f = \{v \in L^2(\Sigma) \mid v(x, t) &\geqq 0 \text{ if } (\Pi f)(x, t) = f(x, t) = 0, \\ v(x, t) &\leqq 0 \text{ if } (\Pi f)(x, t) = f(x, t) = M, \\ v(x, t) &= 0 \text{ if } (\Pi f)(x, t) = 0 \text{ and} f(x, t) < 0 \\ &\text{or } (\Pi f)(x, t) = 0 \text{ and} f(x, t) > M\}. \end{aligned} \tag{2.22}$$

Proof of Lemma 2 is given in [S3].

3. Sensitivity analysis of optimal control problem. For any parameter $\varepsilon \in [0, \delta)$ the solution of state equation (1.3)-(1.5) takes the form (2.4) i.e.,

$$y(t) = (L_\varepsilon u)(t) + C_\varepsilon(t) y_0 + S_\varepsilon(t) y_1, \qquad t \in [0, T]; \tag{3.1}$$

therefore we denote

$$y(T) = L_{T,\varepsilon} u + C_\varepsilon(T) y_0 + S_\varepsilon(T) y_1. \tag{3.2}$$

Cost functional (1.1) can be rewritten as

$$\begin{aligned} I_\varepsilon(u) &= \tfrac{1}{2}(R L_{T,\varepsilon} u, L_{T,\varepsilon} u)_\Omega + (R L_{T,\varepsilon} u, C_\varepsilon(T) y_0 + S_\varepsilon(T) y_1)_\Omega \\ &\quad + \tfrac{1}{2}|u|_\Sigma^2 + \text{const}\,(y_0, y_1) = \tfrac{1}{2}(|u|_\Sigma^2 + (B_\varepsilon u, u)_\Sigma) - (f_\varepsilon, u)_\Sigma + \text{const}, \end{aligned} \tag{3.3}$$

where

$$B_\varepsilon = L^*_{T,\varepsilon} R L_{T,\varepsilon}, \tag{3.4}$$

$$f_\varepsilon = -L^*_{T,\varepsilon} R[C_\varepsilon(T)y_0 + S_\varepsilon(T)y_1]. \tag{3.5}$$

The unique optimal control $u_\varepsilon \in K$ that minimizes functional (3.3) over the set $K \subset L^2(\Sigma)$ can be characterized as the unique fixed point

$$u_\varepsilon = \Pi[-B_\varepsilon u_\varepsilon + f_\varepsilon] \tag{3.6}$$

where the metric projection Π in $L^2(\Sigma)$ onto K is defined by (2.17).

Below we formulate several lemmas that are fundamental for the proof of the main result. The proofs of the lemmas are relegated to § 4.

LEMMA 3 (regularity of optimal controls). *For* $\varepsilon \in [0, \delta)$

$$\|u_\varepsilon\|_{H^1(\Sigma)} \leqq C,$$

where the constant C *depends only on the data of the problem and not on* $\varepsilon > 0$.

From Lemma 1(iii) and Lemma 3 we obtain Corollary 1.

COROLLARY 1. *For* $\rho > 0$, ρ *arbitrary small*

$$Lu_\varepsilon \in C[0, T; H^{3/2-2\rho}(\Omega)]$$

uniformly in $\varepsilon > 0$.

LEMMA 4. *For any* $u \in H^1(\Sigma)$ *we have for* $\varepsilon > 0$, ε *small enough*

$$B_\varepsilon u = B_0 u + \varepsilon B' u + o(\varepsilon) u \quad \text{in } L^2(\Sigma),$$

where

$$B' \in \mathscr{L}(H^1(\Sigma); L^2(\Sigma))$$

is given by (4.18) *and*

$$\forall u \in H^1(\Sigma): \quad |o(\varepsilon)u|_\Sigma / \varepsilon \to 0 \quad \text{with } \varepsilon \downarrow 0.$$

LEMMA 5. *Assume that* $y_0 \in H^{1/2}(\Omega)$, $y_1 \in H^{-1/2}(\Omega)$. *For* $\varepsilon > 0$, ε *small enough*

$$f_\varepsilon = f_0 + \varepsilon f' + o(\varepsilon) \quad \text{in } L^2(\Sigma),$$

where f' *is given by* (4.27), *and* $|o(\varepsilon)|_\Sigma / \varepsilon \to 0$ *with* $\varepsilon \downarrow 0$.

Next we show that u_ε is Lipschitz continuous with respect to ε. Equation (3.6) is equivalent to the following variational inequality.

Find

$$u_\varepsilon \in K: \quad (u_\varepsilon + B_\varepsilon u_\varepsilon - f_\varepsilon, v - u_\varepsilon)_\Sigma \geqq 0 \quad \forall v \in K, \tag{3.7}$$

hence we have

$$(u_\varepsilon + B_\varepsilon u_\varepsilon - f_\varepsilon, u_0 - u_\varepsilon)_\Sigma \geqq 0, \tag{3.8}$$

$$(-u_0 + B_0 u_0 - f_0, u_0 - u_\varepsilon)_\Sigma \geqq 0, \tag{3.9}$$

and therefore

$$\begin{aligned} &|u_0 - u_\varepsilon|^2_\Sigma + (B_\varepsilon(u_0 - u_\varepsilon), u_0 - u_\varepsilon)_\Sigma \\ &\qquad \leqq (f_0 - f_\varepsilon, u_0 - u_\varepsilon)_\Sigma + ((B_0 - B_\varepsilon)u_0, u_0 - u_\varepsilon)_\Sigma. \end{aligned} \tag{3.10}$$

By Lemma 4, it follows that

$$(B_0 - B_\varepsilon)u_0 = \varepsilon B' u_0 + o(\varepsilon) \quad \text{in } L^2(\Sigma), \tag{3.11}$$

where $|o(\varepsilon)|_\Sigma / \varepsilon \to 0$ with $\varepsilon \downarrow 0$.

In view of (3.11) we have

$$|(B_0 - B_\varepsilon)u_0|_\Sigma \leqq C\varepsilon \tag{3.12}$$

for some constant C, which is independent of $\varepsilon \in [0, \delta)$.

Similarly, by Lemma 5

$$|f_0 - f_\varepsilon|_\Sigma \leqq C\varepsilon.$$

Finally from (3.10), in view of (3.12) and since

$$(B_\varepsilon u, u)_\Sigma = (RL_{T,\varepsilon}u, L_{T,\varepsilon}u)_\Sigma \geqq 0 \quad \forall u \in L^2(\Sigma),$$

it follows that

$$|u_\varepsilon - u_0|_\Sigma \leqq C\varepsilon. \tag{3.13}$$

Now we are in a position to prove the theorem.

Proof of Theorem 1. Let ε_n be a sequence such that $\varepsilon_n \downarrow 0$ with $n \to \infty$. In view of (3.13) it follows that

$$\left|\frac{1}{\varepsilon_n}(u_{\varepsilon_n} - u_0)\right|_\Sigma \leqq C,$$

hence there exists an element $q \in L^2(\Sigma)$ such that

$$u_{\varepsilon_n} = u_0 + \varepsilon_n q + r(\varepsilon_n) \quad \text{in } L^2(\Sigma), \tag{3.14}$$

where

$$r(\varepsilon_n)/\varepsilon_n \to 0 \quad \text{with } \varepsilon_n \downarrow 0 \quad \text{weakly in } L^2(\Sigma).$$

In the sequel we denote ε for ε_n.

We will prove that the element q is uniquely determined and characterized as the solution to Problem (Q) and that actually $r(\varepsilon)/\varepsilon \to 0$ strongly in $L^2(\Sigma)$. To this end we shall use the results of Lemmas 3–5.

By Lemmas 4 and 5 we have

$$B_\varepsilon = B_0 + \varepsilon B' + o(\varepsilon) \quad \text{in } \mathscr{L}(H^1(\Sigma); L^2(\Sigma)), \tag{3.15}$$

$$f_\varepsilon = f_0 + \varepsilon f' + o(\varepsilon) \quad \text{in } L^2(\Sigma), \tag{3.16}$$

where

$$\forall u \in H^1(\Sigma): \quad |o(\varepsilon)u|_\Sigma/\varepsilon \to 0 \quad \text{with } \varepsilon \downarrow 0 \tag{3.17}$$

and

$$B_\varepsilon, B_0 \in \mathscr{L}(L^2(\Sigma)) \quad \text{uniformly in } \varepsilon > 0, \tag{3.18}$$

$$B' \in \mathscr{L}(H^1(\Sigma); L^2(\Sigma)), \tag{3.19}$$

and therefore

$$B_\varepsilon u_\varepsilon - f_\varepsilon = B_0 u_0 - f_0 + \varepsilon[B'u_0 + B_0 q - f'] + o(\varepsilon) \quad \text{in } L^2(\Sigma). \tag{3.20}$$

The explicit forms of B', f' are given in (4.18) and (4.27), respectively. From the right-hand side of (3.6), in view of (2.21) and (3.20) it follows that

$$\begin{aligned} u_\varepsilon = \Pi[-B_\varepsilon u_\varepsilon - f_\varepsilon] &= \Pi[-B_0 u_0 + f_0] \\ &+ \varepsilon \Pi'_{\{-B_0 u_0 + f_0\}}[-B'u_0 - B_0 q + f'] + o(\varepsilon) \quad \text{in } L^2(\Sigma), \end{aligned} \tag{3.21}$$

and therefore in view of (3.14) we obtain

$$q = \Pi'_{\{-B_0 u_0 + f_0\}}[-B'u_0 - B_0 q + f'], \tag{3.22}$$

$$r(\varepsilon) = o(\varepsilon). \tag{3.23}$$

Hence by (3.22) we have

$$q \in S_{\{-B_0u_0+f_0\}},$$

since $\Pi': L^2(\Sigma) \to S_{\{-B_0u_0+f_0\}} \subset L^2(\Sigma)$ is the metric projection in $L^2(\Sigma)$ onto the cone $S_{\{-B_0u_0+f_0\}} \subset L^2(\Sigma)$.

Since the optimality conditions for the control problem under consideration read

$$u_0 \in K: \quad \int_\Sigma (u_0 + a_0p_0)(v-u_0)\,d\Sigma$$
$$= (u_0 + B_0u_0 - f_0, v-u_0)_\Sigma \leqq 0 \quad \forall v \in K,$$

it follows that we can identify

$$a_0p_{0|\Sigma} = B_0u_0 - f_0. \tag{3.24}$$

Thus in view of (3.7), it follows that the cone $S_{\{-B_0u_0+f_0\}} = S_{\{-a_0p_{0|\Sigma}\}}$ coincides with

$$\begin{aligned} S \equiv \{v \in L^2(\Sigma) \mid & v(x,t) \geqq 0 \text{ if } u_0(x,t) = -a_0(x)p_0(x,t) = 0, \\ & v(x,t) \leqq 0 \text{ if } u_0(x,t) = -a_0(x)p_0(x,t) = M, \\ & v(x,t) = 0 \text{ if } u_0(x,t) = 0 \text{ and } p_0(x,t) > 0 \\ & \text{or } u_0(x,t) = M \text{ and } a_0(x)p_0(x,t) < -M\}, \end{aligned} \tag{3.25}$$

which, in turn, by (3.7) is equivalent to the cone S defined by (1.10).

By standard arguments in optimization theory, we can show that in view of (1.14)-(1.20) Problem (Q) is equivalent to the following optimality system.

Find (q, z, w) such that the following system is satisfied:

State equation:

$$z_{tt} = A_0(x,\partial)z + \operatorname{div}(a_0\nabla y^0) \quad \text{in } Q, \tag{3.26}$$

$$\frac{\partial z}{\partial n} = q + \frac{a_1}{a_0}u_0 \quad \text{on } \Sigma, \tag{3.27}$$

$$z(0) = 0, \quad z_t(0) = 0 \quad \text{in } \Omega. \tag{3.28}$$

Adjoint state equation:

$$w_{tt} = A_\varepsilon(x,\partial)w + \operatorname{div}(a_0\nabla p^0) \quad \text{in } Q, \tag{3.29}$$

$$\frac{\partial w}{\partial n} = 0 \quad \text{on } \Sigma, \tag{3.30}$$

$$w(T) = 0, \quad w_t(T) = Rz(T) \quad \text{in } \Omega. \tag{3.31}$$

Optimality conditions:

$$q \in S: \quad \int_\Sigma (q + a_0w + a_1p_0)(v-q)\,d\Sigma \geqq 0 \quad \forall v \in S; \tag{3.32}$$

therefore the solution of Problem (Q) takes the following form:

$$q = \Pi'_{\{-a_0p_{0|\Sigma}\}}(-a_0w_{|\Sigma} - a_1p_{0|\Sigma}). \tag{3.33}$$

On the other hand, by using the definitions (3.4), (4.18), and (4.27) of B_0, B', and f', it is straightforward to verify that

$$-a_0w_{|\Sigma} - a_1p_{0|\Sigma} \equiv -B'u_0 + f' \quad \text{in } L^2(\Sigma); \tag{3.34}$$

therefore q given by (3.22) coincides with q defined by (3.33), and thus q defined by (3.22) is the unique solution of Problem (Q).

The proof of the theorem is thus completed. It remains to establish the validity of Lemmas 3-5.

4. Sensitivity analysis of the wave equation. We will prove that operators

$$B_\varepsilon \in \mathscr{L}(L^2(\Sigma); L^2(\Sigma))$$

and elements $f_\varepsilon \in L^2(\Sigma)$ are differentiable with respect to the parameter. We recall that

$$B_\varepsilon = L^*_{T,\varepsilon} R L_{T,\varepsilon} \tag{4.1}$$

where (1.7)(i) is satisfied, i.e.,

$$R \in \mathscr{L}(H^s(\Omega)), \qquad 0 \leqq s \leqq \tfrac{1}{2} \tag{4.2}$$

or (1.7)(ii) is satisfied and

$$f_\varepsilon = -L^*_{T,\varepsilon} R[C_\varepsilon(T)y_0 + S_\varepsilon(T)y_1]. \tag{4.3}$$

We start with the proof of Lemma 3.

Proof of Lemma 3.

Step 1.

$$f_\varepsilon \in H^1(\Sigma) \quad \text{with the norm uniform in } \varepsilon. \tag{4.4}$$

Note first that by (1.7)

$$R[C_\varepsilon(T)y_0 + S_\varepsilon(T)y_1] \in H^{1/2-2\rho}(\Omega) = D(A^{1/4-\rho}). \tag{4.5}$$

Here we have used the regularity of map R together with the well-known properties of sine $S(t)$ and cosine $C(t)$ operators (see [F1]):

$$\begin{aligned} &C(\cdot): D(A^s) \to C[0, T; D(A^s)], \\ &S(\cdot): D(A^s) \to C[0, T; D(A^{s+1/2})]. \end{aligned} \tag{4.6}$$

By Lemma 1(ii) applied to the representation (4.3) and by (4.5), we obtain

$$f_\varepsilon \in H^{1+2\alpha-2\rho}(\Sigma) \quad \text{uniformly in } \varepsilon > 0,$$

which in particular implies (4.4).

Step 2.

$$u_\varepsilon \in H^{4\alpha}(\Sigma) \quad \text{with the norm uniform in } \varepsilon > 0. \tag{4.7}$$

By Lemma 1(i) we have

$$L_{T,\varepsilon} \in \mathscr{L}(L^2(\Sigma); D(A^{1/4+\alpha}) = H^{1/2+2\alpha}(\Omega)).$$

Hence

$$RL_{T,\varepsilon} \in \mathscr{L}(L^2(\Sigma); H^{-1/2+2\alpha}(\Omega) = D(A^{1/4-\alpha})') \tag{4.8}$$

and by interpolating the regularity results given by Lemma 1(ii) we obtain

$$L^*_{T,\varepsilon} \in \mathscr{L}(D(A^{1/4-\alpha})'; H^{4\alpha}(\Sigma)). \tag{4.9}$$

Thus

$$L^*_{T,\varepsilon} R L_{T,\varepsilon} \in \mathscr{L}(L^2(\Sigma); H^{4\alpha}(\Sigma)) \tag{4.10}$$

and by the result of Step 1

$$B_\varepsilon u_\varepsilon + f_\varepsilon \in H^{4\alpha}(\Sigma) \tag{4.11}$$

with the norm uniform in $\varepsilon > 0$. The result in (4.7) follows now from (4.11), (3.6), and (2.20).

Step 3.

$$u_\varepsilon \in H^{8\alpha}(\Sigma). \tag{4.12}$$

By interpolating between the results in Lemma 1(i), we obtain

$$L_{T,\varepsilon} \in \mathscr{L}(H^{4\alpha}(\Sigma); D(A^{1/4+3\alpha}) = H^{1/2+6\alpha}(\Omega)). \tag{4.13}$$

Hence

$$RL_{T,\varepsilon} \in \mathscr{L}(H^{4\alpha}(\Sigma); H^{-1/2+6\alpha}(\Omega) = (D(A^{1/4-3\alpha}))'). \tag{4.14}$$

Interpolating between the results of Lemma 1(ii) gives

$$L^*_{T,\varepsilon} \in \mathscr{L}((D(A^{1/4-3\alpha}))'; H^{8\alpha}(\Sigma)). \tag{4.15}$$

By (4.14), (4.15), and (4.7)

$$L^*_{T,\varepsilon} RL_{T,\varepsilon} u_\varepsilon \in H^{8\alpha}(\Sigma),$$

which result combined with (4.4) gives

$$-B_\varepsilon u_\varepsilon + f_\varepsilon \in H^{8\alpha}(\Sigma).$$

The final conclusion in (4.7) is obtained, as before, by combining (4.15), (3.6), and (2.20).

Step 4.

$$u_\varepsilon \in H^1(\Sigma). \tag{4.16}$$

We apply the "bootstrap" argument by "boosting" successively the regularity of u_ε by the "4α" derivative. The procedure ends when the maximal regularity of the map $L_{T,\varepsilon}$ is achieved. Indeed, at the last step we have

$$L_T \in \mathscr{L}(H^{1-2\alpha-2\rho}(\Sigma); D(A^{3/4-\rho}))$$

and to obtain higher regularity of the map L_T, we need the compatibility condition $u_\varepsilon(t=0)=0$. Since such a condition does not hold, in general, we must stop our argument, obtaining

$$L_T u_\varepsilon \in D(A^{3/4-\rho}) = H^{3/2-2\rho}(\Omega).$$

Applying once more the regularity of $L^*_{T,\varepsilon} R$, we obtain

$$L^*_{T,\varepsilon} RL_T u_\varepsilon \in H^{1+2\alpha-2\rho}(\Sigma) \subset H^1(\Sigma)$$

for ρ arbitrarily small. Hence

$$-B_\varepsilon u_\varepsilon + f_\varepsilon \in H^1(\Sigma)$$

and the application of (2.20) yields the desired conclusion. □

Proof of Lemma 4 will follow through the sequence of propositions whose proofs are given in § 5.

PROPOSITION 1. *Let ρ be arbitrary small. For any $u \in H^{1-2\alpha-2\rho}(\Sigma)$*

$$(L_T - L_{T,\varepsilon})u = \varepsilon L'_T u + r_1(\varepsilon),$$

where

$$\frac{r_1(\varepsilon)}{\varepsilon} \to 0 \quad \textit{in } D(A^{1/4-\rho}) \quad \textit{with } \varepsilon \to 0$$

and

$$L'_T \in \mathscr{L}(H^{1-2\alpha-2\rho}(\Sigma);\ D(A^{1/4-\rho})).$$

PROPOSITION 2. *For any* $u \in H^{1-2\alpha-2\rho}(\Sigma)$

$$L_T^* R(L_T - L_{T,\varepsilon})u = \varepsilon L_T^* R L'_T u + r_2(\varepsilon)$$

where

$$\frac{r_2(\varepsilon)}{\varepsilon} \to 0 \quad \textit{in } L^2(\Sigma) \quad \textit{with } \varepsilon \to 0$$

and

$$L_T^* R L'_T \in \mathscr{L}(H^{1-2\alpha-2\rho}(\Sigma);\ L^2(\Sigma)).$$

PROPOSITION 3. *For any* $f \in D(A^{1/4-\rho})$

$$(L_T^* - L_{T,\varepsilon}^*)f = \varepsilon L_T^{*\prime} f + r_3(\varepsilon)$$

where

$$\frac{r_3(\varepsilon)}{\varepsilon} \to 0 \quad \textit{in } L^2(\Sigma) \quad \textit{with } \varepsilon \to 0$$

and

$$L_T^{*\prime} \in \mathscr{L}(D(A^{1/4-\rho});\ L^2(\Sigma)) \quad \textit{and} \quad L_T^{*\prime} \quad \textit{is given by } (5.23).$$

PROPOSITION 4. *For any* $u \in H^{1-2\alpha-\rho}(\Sigma)$

$$(L_T^* - L_{T,\varepsilon}^*) R L_T u = \varepsilon L_T^{*\prime} R L_T u + r_4(\varepsilon)$$

where

$$\frac{r_4(\varepsilon)}{\varepsilon} \to 0 \quad \textit{in } L^2(\Sigma) \quad \textit{with } \varepsilon \to 0$$

and

$$L_T^{*\prime} R L_T \in \mathscr{L}(H^{1-2\alpha-2\rho}(\Sigma);\ L^2(\Sigma)).$$

Proof of Lemma 4. We have

$$\begin{aligned}(B_0 - B_\varepsilon)u_0 &= (L_T^* - L_{T,\varepsilon}^*) R L_T u_0 + L_T^* R(L_T - L_{T,\varepsilon})u_0 + (L_{T,\varepsilon}^* - L_T^*) R(L_T - L_{T,\varepsilon})u_0 \\ &= \varepsilon[L_T^{*\prime} R L_T + L_T^* R L'_T]u_0 + o(\varepsilon) \\ &= \varepsilon B' u_0 + o(\varepsilon),\end{aligned}$$

where

(4.17) $$o(\varepsilon) = r_2(\varepsilon) + r_4(\varepsilon) - (L_{T,\varepsilon}^* - L_T)R(L_T - -L_{T,\varepsilon})u_0,$$

(4.18) $$B'u \equiv (L_T^{*\prime} R L_T + L_T^* R L'_T)u.$$

From Propositions 2 and 4 it follows that with L'_T and $L_T^{*\prime}$ given by (5.4) and (5.23)

$$B' \in \mathscr{L}(H^{1-2\alpha-2\rho}(\Sigma);\ L^2(\Sigma)),$$

hence in particular

$$B' \in \mathscr{L}(H^1(\Sigma);\ L^2(\Sigma))$$

and $r_2(\varepsilon)/\varepsilon \to 0$, $r_4(\varepsilon)/\varepsilon \to 0$ in $L^2(\Sigma)$ with $\varepsilon \to 0$. We also have

$$\frac{1}{\varepsilon}(L^*_{T,\varepsilon} - L^*_T)R(L_T - L_{T,\varepsilon})u_0 = (L^*_{T,\varepsilon} - L^*_T)R\frac{1}{\varepsilon}[\varepsilon L'_T u_0 + r(\varepsilon)].$$

By Proposition 1

(4.19) $R(L'_T u_0 + r_1(\varepsilon)/\varepsilon)$ is uniformly bounded in $H^{-1/2-2\rho}(\Omega) = D(A^{1/4+\rho})'$.

Moreover,

$$r_1(\varepsilon)/\varepsilon \to 0 \quad \text{in } D(A^{1/4+\rho})'. \tag{4.20}$$

On the other hand, for all $f \in D(A^{1/4+\rho})'$

$$\|(L^*_{T,\varepsilon} - L^*_T)f\|_{L^2(\Sigma)} \to 0. \tag{4.21}$$

Indeed from Lemma 1(ii) we obtain, in particular,

$$\|(L^*_{T,\varepsilon} - L^*_T)f\|_{L^2(\Sigma)} \leqq C_T \|f\|_{(D(A^{1/4+\rho}))'}. \tag{4.22}$$

On the other hand, it can be shown by standard methods that for $f \in D(A^{1/2})$

$$\|(L^*_{T,\varepsilon} - L^*_T)f\|_{L^2(\Sigma)} \to 0. \tag{4.23}$$

To see (4.23) it is enough to write

$$w(t) = (L^*_{T,\varepsilon} - L^*_T)f$$

with

$$\begin{aligned} &w(t) = a_0 \bar{w}(t)_{|\Gamma} + (a_0 - a_\varepsilon)z_{\varepsilon|\Gamma}, \\ &\bar{w}_{tt} - A\bar{w} = \varepsilon \operatorname{div}[(a_1 + r_0/\varepsilon)\nabla S_\varepsilon(T-t)f], \\ &\bar{w}(T) = \bar{w}_t(T) = 0, \qquad \bar{w}_{|\Gamma} = 0, \\ &z_\varepsilon(t) = S_\varepsilon(T-t)f. \end{aligned} \tag{4.24}$$

Since $S_\varepsilon(T-t)f \in C[0, T; D(A) \subset H^2(\Omega)]$ uniformly in ε

$$\|z_\varepsilon(t)_{|\Gamma}\| \leqq C \quad \text{in } C[0, T; H^{3/2}(\Gamma)] \tag{4.25}$$

and the right-hand side of (4.24) converges to zero strongly in $C[0, T; L^2(\Omega)]$. By standard results

$$\bar{w} \to 0 \quad \text{in } C[0, T; H^1(\Omega)],$$

hence, in particular by the trace theorem,

$$w \equiv \bar{w}_{|\Gamma} \to 0 \quad \text{in } L^2(\Sigma).$$

This together with (4.25) proves (4.23). Formulas (4.22), (4.23) and density of $D(A^{1/2})$ in $D(A^{1/4+\rho})'$ imply (4.21). To complete the proof it is enough to note that by (4.19), (4.20), and (4.22)

$$(L^*_{T,\varepsilon} - L^*_T)Rr_1(\varepsilon)/\varepsilon \to 0 \quad \text{in } L^2(\Sigma)$$

and by (4.19), (4.20), and (4.21)

$$(L^*_{T,\varepsilon} - L^*_T)RL'_T u_0 \to 0 \quad \text{in } L^2(\Sigma).$$

Hence

$$(L^*_{T,\varepsilon} - L^*_T)R(L'_T u_0 + r_1(\varepsilon)/\varepsilon) \to 0 \quad \text{in } L^2(\Sigma),$$

which completes the proof of Lemma 4. □

Proof of Lemma 5. We assume that

$$y_0 \in D(A^{1/4}), \qquad y_1 \in (D(A^{1/4}))'$$

and we denote

$$\begin{aligned} X(t) &\equiv S(t)y_1 + C(t)y_0, \\ X_\varepsilon(t) &\equiv S_\varepsilon(t)y_1 + C_\varepsilon(t)y_0, \\ w(t) &\equiv X(t) - X_\varepsilon(t). \end{aligned}$$

Then

$$\begin{aligned} &w_{tt} = Aw(t) + (A - A_\varepsilon)X_\varepsilon(t) \quad \text{in } Q, \\ &\frac{\partial w}{\partial n} = 0 \quad \text{on } \Sigma, \\ &w(0) = w_t(0) = 0 \quad \text{in } \Omega. \end{aligned}$$

Let $a_\varepsilon = a_0 + \varepsilon a_1 + r_0(\varepsilon)$.

Since

$$\begin{aligned} (A - A_\varepsilon)y &= -\varepsilon \operatorname{div}\left[\left(a_1 + \frac{r_0(\varepsilon)}{\varepsilon}\right)\nabla y\right], \\ (A - A_\varepsilon)X_\varepsilon &= -\varepsilon[g_\varepsilon \ddot{X}_\varepsilon + a_\varepsilon \nabla g_\varepsilon \cdot \nabla X_\varepsilon], \end{aligned}$$

where

$$g_\varepsilon \equiv \left(a_1 + \frac{r_0(\varepsilon)}{\varepsilon}\right)\frac{1}{a_\varepsilon} \to \frac{a_1}{a_0} \equiv g \quad \text{with } \varepsilon \downarrow 0.$$

Hence

$$\begin{aligned} w(t) &= \int_0^t S(t-\tau)(A - A_\varepsilon)X_\varepsilon(\tau)\,d\tau \\ &= -\varepsilon \int_0^t S(t-\tau)\left[g_\varepsilon \frac{d^2}{d\tau^2}X_\varepsilon(\tau) + a_\varepsilon \nabla g_\varepsilon \cdot \nabla X_\varepsilon(\tau)\right] d\tau. \end{aligned}$$

Therefore

$$\begin{aligned} w(t) &= -\varepsilon \int_0^t S(t-\tau)\left[g \frac{d^2}{d\tau^2}X(\tau) + a_0 \nabla g \cdot \nabla X(\tau)\right] d\tau + r(\varepsilon) \\ &= \varepsilon X'(y_0, y_1) + r(\varepsilon), \end{aligned}$$

where

$$\begin{aligned} \frac{1}{\varepsilon} r(\varepsilon) = \int_0^t S(t-\tau)\Bigg\{&\left[g_\varepsilon \frac{d^2}{d\tau^2}X_\varepsilon(\tau) - g\frac{d^2}{d\tau^2}X(\tau)\right] \\ &+ [a_0 \nabla g \cdot \nabla X(\tau) - a_\varepsilon \nabla g_\varepsilon \cdot \nabla X_\varepsilon(\tau)]\Bigg\}\,d\tau. \end{aligned}$$

Let us consider the case (1.7)(ii) (case (1.7)(i) is simpler, and hence is omitted). We shall show that

$$X'(y_0, y_1) \equiv -\int_0^t S(t-\tau)\left[g\frac{d^2}{d\tau^2}X(\tau) + a_0 \nabla\left(\frac{a_1}{a_0}\right) \cdot \nabla X(\tau)\right] d\tau \tag{4.26}$$

belongs to $C[0, T; D(A^{1/4-\rho})]$, and $(1/\varepsilon)r(\varepsilon) \to 0$ in $C[0, T; D(A^{1/4-\rho})]$.

In fact, since

$$\frac{d^2x}{dt^2}(\cdot)=AC(\cdot)y_0+AS(\cdot)y_1\in C[0, T; (D(A^{1/4+\rho}))']$$

and

$$\nabla X\in C[0, T; D(A^{1/4+\rho})'],$$

(4.6) implies that

$$X'(y_0, y_1)\in C[0, T; D(A^{1/4-\rho})].$$

Since

$$\|X_\varepsilon - X\|_{C[0,T;D(A^{3/4-\rho})]}\to 0,$$

using a similar argument as before, we can also show that

$$r(\varepsilon)/\varepsilon\to 0 \quad \text{in } C[0, T; D(A^{1/4-\rho})].$$

By (4.26), regularity of R, and Lemma 1(ii), we obtain

$$L_T^*RX'(y_0, y_1)\in L^2(\Sigma).$$

Since

$$R[C_\varepsilon(T)y_0+S_\varepsilon(T)y_1]\in D(A^{1/4-\rho}),$$

Proposition 3 yields

$$L_{T,\varepsilon}^{*\prime}R[C_\varepsilon(T)y_0+S_\varepsilon(T)y_1]\in L^2(\Sigma).$$

Writing

$$f_\varepsilon-f_0=(L_T^*-L_{T,\varepsilon}^*)R[C(T)y_0+S(T)y_1]+L_{T,\varepsilon}^*R[X-X_\varepsilon],$$

and using the result of Proposition 3, Lemma 1(ii), and (4.26), it follows that

$$f_\varepsilon-f_0=\varepsilon f'+r(\varepsilon),$$

where $r(\varepsilon)/\varepsilon\to 0$ in $L^2(\Sigma)$ and

$$f'=L_T^{*\prime}R[C(T)y_0+S(T)y_1]+L_T^*RX'\in L^2(\Sigma) \tag{4.27}$$

with X' given by (4.26) and $L_T^{*\prime}$ given by Proposition 3 (see also (5.23)). This completes the proof of Lemma 5.

5. Proofs of Propositions 1–4.

Proof of Proposition 1. Let

$$w\equiv(L-L_\varepsilon)u.$$

Then

$$\begin{aligned} &w_{tt}=Aw+(A(x,\partial)-A_\varepsilon(x,\partial))L_\varepsilon u \quad \text{in } Q,\\ &\frac{\partial w}{\partial n}=0 \quad \text{on } \Sigma,\\ &w(0)=w_t(0)=0 \quad \text{in } \Omega. \end{aligned} \tag{5.1}$$

With

$$a_\varepsilon=a_0+\varepsilon a_1+r_0(\varepsilon)$$

we have

$$|A(x, \partial) - A_\varepsilon(x, \partial)]y = -\varepsilon \operatorname{div}\left[\left(a_1 + \frac{r_0(\varepsilon)}{\varepsilon}\right)\nabla y\right],$$

hence

$$|A(x, \partial) - A_\varepsilon(x, \partial)]L_\varepsilon u = -\varepsilon\left[g_\varepsilon \frac{d^2}{dt^2} L_\varepsilon u + a_\varepsilon \nabla g_\varepsilon \cdot \nabla L_\varepsilon u\right],$$

where

$$g_\varepsilon \equiv \left[a_1 + \frac{r_0(\varepsilon)}{\varepsilon}\right]\frac{1}{a_\varepsilon} \to \frac{a_1}{a_0} \equiv g \quad \text{with } \varepsilon \downarrow 0. \tag{5.2}$$

The convergence takes place in $C^1(\bar{\Omega})$. Thus after denoting

$$\mathbf{h}_\varepsilon \equiv \nabla g_\varepsilon a_\varepsilon \to \nabla\left(\frac{a_1}{a_0}\right) a_0 \equiv \mathbf{h}, \tag{5.3}$$

$$(L_T - L_{T,\varepsilon})u = w(T) = -\varepsilon \int_0^T S(T-\tau)\left[g_\varepsilon \frac{d^2}{d\tau^2}(L_\varepsilon u)(\tau) + \mathbf{h}_\varepsilon \cdot \nabla(L_\varepsilon u)\, d\tau\right.$$

$$= \varepsilon \int_0^T S(T-\tau)\left[g \frac{d^2}{d\tau^2}(Lu)(\tau) + \mathbf{h} \cdot \nabla(Lu)(\tau)\right] d\tau - \varepsilon \int_0^T S(T-\tau)$$

$$\cdot \left[g_\varepsilon \frac{d^2}{d\tau^2}(Lu)(\tau) - g\frac{d^2}{d\tau^2}(Lu)(\tau)) + (\mathbf{h}_\varepsilon \cdot \nabla(L_\varepsilon u)(\tau) - \mathbf{h} \cdot \nabla(Lu)(\tau))\right] d\tau.$$

Let us denote

$$\begin{aligned} L'_T u &\equiv -\int_0^T S(T-\tau)\left[g\frac{d^2}{d\tau^2}(Lu)(\tau) + \mathbf{h}\cdot\nabla(Lu)(\tau)\right] d\tau \\ &= -\int_0^T S(T-\tau) \operatorname{div}(a_1 \nabla Lu(\tau))\, d\tau, \end{aligned} \tag{5.4}$$

$$\begin{aligned} -r_1(\varepsilon, u)/\varepsilon \equiv &\int_0^T S(T-\tau)\left[g_\varepsilon \frac{d^2}{d\tau^2}(L_\varepsilon u)(\tau) - g\frac{d^2}{d\tau^2}(Lu)(\tau)\right] d\tau \\ &+ \int_0^T S(T-\tau)[\mathbf{h}_\varepsilon \cdot \nabla(L_\varepsilon u(\tau)) - \mathbf{h}\cdot\nabla(Lu)(\tau)]\, d\tau. \end{aligned} \tag{5.5}$$

To prove Proposition 1 we need to show that

$$L'_T \in \mathscr{L}(H^{1-2\alpha-2\rho}(\Sigma); D(A^{1/4-\rho})), \tag{5.6}$$

$$r_1(\varepsilon, u)/\varepsilon \to 0 \quad \text{in } D(A^{1/4-\rho}), \quad u \in H^{1-2\alpha-2\rho}(\Sigma). \tag{5.7}$$

To accomplish this we first observe that

$$\frac{d^2}{dt^2} L_\varepsilon \in \mathscr{L}(H^{1-2\alpha-2\rho}(\Sigma); C[0, T; (D(A^{1/4+\rho}))']), \tag{5.8}$$

$$\nabla L_\varepsilon \in \mathscr{L}(H^{1-2\alpha-2\rho}(\Sigma); C[0, T; D(A^{1/4-\rho})]). \tag{5.9}$$

In fact from Lemma 1(i), we obtain

$$AL \in \mathscr{L}(H^{1-2\alpha-2\rho}(\Sigma); C[0, T; (D(A^{1/4+\rho}))']).$$

We also have that

$$ANLu = A^{1/4+\rho}A^{3/4-\rho}N, \qquad Lu \in (D(A^{1/4+\rho}))' \text{ bounded.}$$

Thus with $u \in H^{1-2\alpha-2\rho}(\Sigma)$

$$ALu - ANLu \in C[0, T; (D(A^{1/4+\rho}))']. \tag{5.10}$$

On the other hand, we have

$$\frac{d^2}{dt^2}(Lu) = [A - AN]Lu. \tag{5.11}$$

Therefore (5.8) follows from (5.10), and (5.11) and (5.9) follow from Lemma 1(ii) and from the derivative theorem in [LM]. To continue with the proof of (5.6) we recall that for any $\gamma > 0$

$$S(\cdot) \in \mathscr{L}(D(A^\gamma); C[0, T; D(A^{\gamma+1/2})]), \tag{5.12}$$

$$C(\cdot) \in \mathscr{L}(D(A^\gamma); C[0, T; D(A^\gamma)]). \tag{5.13}$$

Thus by the virtue of (5.8) the expression in the bracket in (5.4) is in $C[0, T; (D(A^{1/4-\rho}))']$ for $u \in H^{1-2\alpha-2\rho}(\Sigma)$, hence by (5.12), (5.13) $L'_T u \in D(A^{1/4-\rho})$ for $u \in H^{1-2\alpha-2\rho}(\Sigma)$ which a posteriori implies (5.6). As for (5.7), we note that by the same argument as before we have

$$\|r_1(\varepsilon, u)/\varepsilon\|_{D(A^{1/4-\rho})} \leqq C\|u\|_{H^{1-2\alpha-2\rho}(\Sigma)}. \tag{5.14}$$

Thus to prove (5.7), we first show that

$$\|L_\varepsilon u - Lu\|_{L^2[0,T;D(A^{3/4-\rho})]} \to 0 \quad \text{for } u \in H^{1-2\alpha-2\rho}(\Sigma). \tag{5.15}$$

Indeed, from (5.1)-(5.3)

$$(L_\varepsilon u - Lu)(t) = w(t) = -\varepsilon \int_0^t S(t-\tau)\left[g_\varepsilon \frac{d^2}{d\tau^2}(L_\varepsilon u)(\tau) + \mathbf{h}_\varepsilon \cdot \nabla(Lu)(\tau)\right] d\tau. \tag{5.16}$$

By (5.2), (5.3), (5.8), (5.9), and (5.12)

$$\|L_\varepsilon u - Lu\|_{C[0,T;D(A^{1/4-\rho})]} \to 0. \tag{5.17}$$

On the other hand, from Lemma 1(iv) we have

$$\|L_\varepsilon u - Lu\|_{H^{3/2-\rho}(Q)} \leqq C \quad \text{uniformly in } \varepsilon > 0. \tag{5.18}$$

By compactness of the imbedding

$$H^{3/2-\rho}(Q) \subset H^{3/2-2\rho}(Q)$$

and by (5.17), we obtain

$$\|L_\varepsilon u - Lu\|_{H^{3/2-2\rho}(Q)} \to 0, \tag{5.19}$$

which implies (5.15). The next step is to prove that for $u \in H^{1-2\alpha-2\rho}(\Sigma)$,

$$\frac{r_1(\varepsilon, u)}{\varepsilon} \to 0 \quad \text{in } D(A^{1/4-\rho}) \quad \text{with } \varepsilon \to 0.$$

Indeed, this follows from the representation (5.5), (5.12) and from the following convergence results which, in turn, are the consequences of (5.10), (5.11), (5.15):

$$\|\nabla(L_\varepsilon - L)u\|_{L^2[0,T;D(A^{1/4-\rho})]} \to 0, \tag{5.20}$$

$$\left\|\frac{d^2}{dt^2}(L_\varepsilon - L)u\right\|_{L^2[0,T;(D(A^{1/4+\rho}))']} \to 0. \tag{5.21}$$

□

Proof of Proposition 2. This follows from Proposition 1, from the assumption (1.7) which implies

$$R \in \mathscr{L}(H^{1/2-\rho}(\Omega); (H^{1/2+\rho}(\Omega))'),$$

and from smoothing property of L_T^*, i.e.,

$$L_T^* \in \mathscr{L}((D(A^{1/4+\alpha}))'; L^2(\Sigma))$$

(see Lemma 1(ii)). □

Proof of Proposition 3. Let

$$w(t) \equiv [S(T-t) - S_\varepsilon(T-t)]f.$$

Then

$$(L_T^* - L_{T,\varepsilon}^*)f = [a_0 S(T-t) - a_\varepsilon S_\varepsilon(T-t)]f_{|\Gamma} = a_0 w_{|\Gamma} + (a_0 - a_\varepsilon) z_{|\Gamma},$$

where $w(t)$ satisfies

$$\begin{aligned} &w_{tt} = Aw + (A - A_\varepsilon) S_\varepsilon(T-t) \quad \text{in } Q, \\ &\frac{\partial w}{\partial n} = 0 \quad \text{on } \Sigma, \qquad (5.22) \\ &w(T) = w_t(T) = 0 \quad \text{in } \Omega \end{aligned}$$

and $z(t) = S_\varepsilon(T-t)f$. Here (see (5.2), (5.3))

$$(A - A_\varepsilon) S_\varepsilon(T-t)f = -\varepsilon \left[g_\varepsilon \frac{d^2}{dt^2} S_\varepsilon(T-t)f + \mathbf{h}_\varepsilon \cdot \nabla S_\varepsilon(T-t)f \right].$$

Thus

$$\begin{aligned} w(t) &= \int_t^T S(\tau - t)(A - A_\varepsilon) S_\varepsilon(T-\tau) f\, dt \\ &= \varepsilon \int_t^T S(\tau - t)\left[g \frac{d^2}{d\tau^2}(S(T-\tau)f) + \mathbf{h} \cdot \nabla(S(T-\tau)f) \right] d\tau \\ &\quad + \varepsilon \int_t^T S(\tau - t) \left[g_\varepsilon \frac{d^2}{d\tau^2}(S_\varepsilon(T-\tau)f) \right. \\ &\quad \left. - g \frac{d^2}{d\tau^2}(S(T-\tau)f) + \mathbf{h}_\varepsilon \cdot \nabla(S_\varepsilon(T-\tau)f) - \mathbf{h} \cdot \nabla(S(T-\tau)f) \right] d\tau. \end{aligned}$$

Denote

$$\begin{aligned} L_T^{*\prime} f &\equiv -L^* \left[g \frac{d^2}{d\tau^2}(S(T-\tau)f) + \mathbf{h} \cdot \nabla(S(T-\tau)f) \right] \\ &= a_1 S(T-\cdot) f_{|\Gamma} \qquad (5.23) \\ &= -L^*[\operatorname{div}(a_1 \nabla(S(T-\cdot)f))] + \frac{a_1}{a_0} L_T^* f, \end{aligned}$$

$$\begin{aligned} r_3(\varepsilon,f)/\varepsilon \equiv L^*\Big[& g_\varepsilon \frac{d^2}{d\tau^2}(S_\varepsilon(T-\tau)f) - g\frac{d^2}{d\tau^2}(S(T-\tau)f) \\ & + \mathbf{h}_\varepsilon \cdot \nabla(S_\varepsilon(T-\tau)f) - \mathbf{h}\cdot\nabla(S(T-\tau)f)\Big] \\ & + \frac{r_1(\varepsilon)}{a_0} L_T^* f + (a_1 + r_1/\varepsilon) w_{|\Gamma}. \end{aligned} \tag{5.24}$$

Then

$$(L_T^* - L_{T,\varepsilon}^*)f = \varepsilon L_T^{*\prime} f + r_3(\varepsilon, f).$$

To prove Proposition 3 we need to show that

$$L_T^{*\prime} \in \mathscr{L}(D(A^{1/4-\rho});\, L^2(\Sigma)), \tag{5.25}$$

$$r_3(\varepsilon,f)/\varepsilon \to 0 \quad \text{in } L^2(\Sigma) \quad \text{for} f \in D(A^{1/4-\rho}). \tag{5.26}$$

Since

$$\frac{d^2}{d\tau^2}(S(T-\tau)f) = AS(T-\tau)f$$

By (5.12), (5.13)

$$\frac{d^2}{dt^2} S(T-\cdot) \in \mathscr{L}(D(A^{1/4-\rho});\, C[0,T;(D(A^{1/4+\rho}))']). \tag{5.27}$$

Also

$$\begin{aligned} \|\nabla(S(t-\cdot)f)\|_{c[0,T;(D(A^{1/4+\rho}))']} & \leqq C\|\nabla(S(T-\cdot)f)\|_{L^2(Q;R^n)} \\ & \leqq C\|A^{1/2}S(T-\cdot)f\|_{L^2(Q)} \leqq C\|f\|_{L^2(\Omega)} \\ & \leqq C\|f\|_{D(A^{1/4-\rho})}. \end{aligned} \tag{5.28}$$

Lemma 1(iii), after using duality, we obtain in particular

$$L^* \in \mathscr{L}(L_1(0,T;\, D(A^{1/4+\alpha})');\, L^2(\Sigma)). \tag{5.29}$$

(5.25) follows now from (5.27), (5.28), Lemma 1(ii), and (5.23).

As for (5.26), we write

$$\frac{r_3(\varepsilon,f)}{\varepsilon} = \frac{\hat{r}_3(\varepsilon,f)}{\varepsilon} + \frac{r_3^*(\varepsilon,f)}{\varepsilon}$$

where

$$\frac{r_3^*(\varepsilon,f)}{\varepsilon} = \frac{r_1(\varepsilon)}{a_0} L_T^* f + \left(a_0 + \frac{r_1}{\varepsilon}\right) w_{|\Gamma}.$$

By the result of Lemma 1(ii)

$$r_1(\varepsilon)\|L_T^* f\|_{L^2(\Sigma)} \to 0, \qquad f \in D(A^{1/4-\rho}).$$

Noting that

$$w_{|\Gamma}=\frac{\varepsilon}{a_0} L_T^*\left[g_\varepsilon \frac{d^2}{dt^2} S_\varepsilon(T-t)f+\mathbf{h}_\varepsilon \nabla(S_\varepsilon(T-t)f)\right]$$

and that

$$\frac{d^2}{dt^2} S_\varepsilon(T-\cdot)f=A_\varepsilon S_\varepsilon(T-\cdot)f\in(D(A^{3/4+\rho}))',$$

$$\nabla S_\varepsilon(T-\cdot)f\in D(A^{1/4-\rho}).$$

By (5.29) we also obtain that

$$\left(a_0+\frac{r_1}{\varepsilon}\right) w_{|\Gamma}\to 0 \quad \text{in } L^2(\Sigma).$$

Thus

$$\frac{r_3^*(\varepsilon,f)}{\varepsilon}\to 0 \quad \text{in } L^2(\Sigma).$$

As for $\hat{r}_3$, note

$$|\hat{r}_3(\varepsilon,f)/\varepsilon|_\Sigma \leqq C\|f\|_{D(A^{1/4-\rho})}.$$

To prove (5.26) it is enough to show that

$$|\hat{r}_3(\varepsilon,f)/\varepsilon|_\Sigma\to 0 \quad \text{for } f\in D \tag{5.30}$$

where D is a dense subset of $D(A^{1/4-\rho})$. In fact, let

$$D\equiv D(A).$$

Then it is straightforward to show that

$$\|[S_\varepsilon(T,\cdot)-S(T,\cdot)]f\|_{C[0,T;D(A)]}\to 0. \tag{5.31}$$

Consequently,

$$\left\|\frac{d^2}{dt^2}[S_\varepsilon(T,\cdot)-S(T,\cdot)]f\right\|_{C[0,T;L^2(\Omega)]}\to 0 \tag{5.32}$$

and

$$\|\nabla[S_\varepsilon(T,\cdot)-S(T,\cdot)]f\|_{C(0,T;L^2(\Omega)]}\to 0. \tag{5.33}$$

Formula (5.2), (5.3), (5.32), and (5.33) applied to the representation (5.24) together with regularity result (5.29) imply (5.26) as desired. □

Proof of Proposition 4. This follows from Proposition 3, after taking into account the regularity of mapping R, as well as the regularity of mapping L_T given in Lemma 1(i), i.e., $L_T u\in D(A^{3/4-\rho})$ (see Corollary 1). □

REFERENCES

[F1] H. O. FATTORINI, *The Cauchy problem*, Encyclopedia of Mathematics and Its Applications, Benjamin and Dummings, Reading, MA, 1983.

[FP] S. FITZPATRICK AND R. R. PHELPS, *Differentiability of the metric projection in Hilbert space*, Trans. Amer. Math. Soc., 207 (1982), pp. 483-501.

[H1] A. HARAUX, *How to differentiate the projection on a convex set in Hilbert space. Some applications to variational inequalities*, J. Math. Soc. Japan, 29 (1977), pp. 615-631.

[HSZ] P. HOLNICKI, J. SOKOŁOWSKI, AND A. ZOCHOWSKI, *Differential stability of solutions to air quality control problems in urban area*, Apl. Mat., 32 (1987), pp. 240-253.

[LT1] I. LASIECKA AND R. TRIGGIANI, *Sharp regularity results for second order hyperbolic equations of Neumann type*, Ann. Mat. Pura Appl., (4), to appear.

[LT2] ———, *A cosine operator approach to modelling $L^2(0, T; L^2(\Gamma))$ boundary input input hyperbolic equations*, Appl. Math. Optim., 7 (1981), pp. 35-93.

[LT3] ———, *Regularity of hyperoblic equations under $L^2(0, T; L^2(\Gamma)$ Dirichlet boundary terms*, Appl. Math. Optim., 10 (1983), pp. 275-286.

[LT4] ———, *Dirichlet boundary control problems for parabolic equations with quadratic cost: analyticity and Riccati's feedback synthesis*, SIAM J. Control Optim., 21 (1984), pp. 41-67.

[LLT] I. LASIECKA, J. L. LIONS, AND R. TRIGGIANI, *Nonhomogeneous boundary value problems for second order hyperbolic operators*, J. Math. Pures Appl., (1986), pp. 149-192.

[L1] I. LASIECKA, *Unified theory for abstract parabolic boundary problems via semigroup approach*, Appl. Math. Optim., 6 (1980), pp. 31-62.

[L2] J. L. LIONS, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, 1971.

[L3] ———, *Perturbations Singulieres dans les Problemes aux Limites et en Controle Optimal*, Lecture Notes in Math., Springer-Verlag, Berlin, 1973.

[LM] J. L. LIONS AND E. MAGENES, *Problemes aux Limites non Homogenes et Applications*, Vol. 1, Dunod, Paris, 1968.

[MS] K. MALANOWSKI AND J. SOKOŁOWSKI, *Sensitivity of solutions to convex, control constrained optimal control problems for distributed parameter systems*, J. Math. Anal. Appl., 120 (1986), pp. 240-263.

[M] F. MIGNOT, *Controle dans les inequations variationelles elliptiques*, J. Funct. Anal., 22 (1976), pp. 25-39.

[N] J. NECAS, *Les Methodes Directes en Theories des Equations*, Masson, Paris, 1967.

[RS1] M. RAO AND J. SOKOŁOWSKI, *Shape sensitivity analysis of state constrained optimal control problems for distributed parameter systems*, Lecture Notes in Control and Information Sciences, Vol. 114, Springer-Verlag, Berlin, New York, 1989, pp. 236-245.

[S1] J. SOKOŁOWSKI, *Sensitivity analysis for a class of variational inequalities*, in Optimization of Distributed Parameter Structures, Vol. 2, E. J. Haug and J. Cea, eds., Sijthoff and Noordhoff, Rockville, MD, 1981, pp. 1600-1609.

[S2] ———, *Differential stability of solutions to constrained optimization problems*, Appl. Math. Optim., 13 (1985), pp. 97-115.

[S3] ———, *Differential Stability of Control Constrained Optimal Control Problems for Distributed Parameter Systems*, Lecture Notes in Control and Information Sciences, Vol. 75, Springer-Verlag, Berlin, New York, 1985, pp. 382-399.

[S4] ———, *Sensitivity analysis and parametric optimization of optimal control problems for distributed parameter systems*, Habilitation Thesis, Warsaw Technical University Publications, Prace Naukowe, Elektronika, z. 73, 1985. (In Polish.)

[S5] ———, *Differential Stability of Solutions to Boundary Optimal Control Problems for Parabolic Systems*, Lecture Notes in Control and Information Sciences, Vol. 84, Springer-Verlag, Berlin, New York, 1986, pp. 854-865.

[S6] ———, *Sensitivity analysis of control constrained optimal control problems for distributed parameter systems*, SIAM J. Control Optim., 25 (1987), pp. 1542-1556.

[S7] ———, *Shape sensitivity analysis of boundary optimal control problems for parabolic systems*, SIAM J. Control Optim., 26 (1988), pp. 763-787.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 5, pp. 1150-1159, September 1991

007

BOUNDARY CONTROLLABILITY OF THE COINCIDENCE SET IN THE OBSTACLE PROBLEM*

VIOREL BARBU† AND DAN TIBA‡

Abstract. Considered is a controlled obstacle problem on a domain Ω with the control $u \in U$ appearing in the boundary data. Given an open and smooth subset D of Ω, it is proved via optimization arguments that the problem $D \subset E_u$ is approximately solvable. Here E_u is the coincidence set corresponding to boundary control u.

Key words. obstacle problem, coincidence set, variational inequality, optimal control

1. Introduction. Let Ω and D be bounded domains of R^N, $N \geqq 2$, such that $D \subset \Omega$. We will assume that $\partial\Omega$ and ∂D are smooth manifolds of class C^m where m is a natural controlled number such that $m > N/2$.

Consider the controlled obstacle problem on

$$\begin{aligned} &\Delta y = 0 \quad \text{in } \{x \in \Omega;\ y(x) > \varphi(x)\},\\ &\Delta y \leqq 0 \quad \text{in } \Omega,\\ &y \geqq \varphi \quad \text{in } \Omega,\\ &y = u \quad \text{in } \partial\Omega, \end{aligned} \tag{1.1}$$

where $u \in H^{m-1/2}(\partial\Omega)$, and $\varphi \in C^m(\bar{\Omega})$.

For any u, we denote by $E_u = \{x \in \Omega;\ y(x) = \varphi(x)\}$ the coincidence set of the obstacle problem and introduce the notation $\Omega_0 = \Omega \backslash \bar{D}$.

In terms of variational inequality, problem (1.1) can be written as

$$y \in K_u;\ \int_\Omega \nabla y \cdot \nabla(y - z)\, dx \leqq 0, \quad \forall z \in K_u, \tag{1.2}$$

where $K_u = \{y \in H^1(\Omega);\ y \geqq \varphi \text{ in } \Omega,\ y = u \text{ in } \partial\Omega\}$.

We will study the following controllability problem:

(1.3) Find $u \in H^{m-1/2}(\partial\Omega)$ such that $D \subset E_u$.

This, in general, is not a well-posed problem. A common way to solve it is to transform it, via the least squares approach, into an optimal control problem governed by the variational inequality (1.2) (see [1], [4]). Here we will approach this problem by reformulating it as follows:

(P) Find $u \in H^{m-1/2}(\partial\Omega)$ such that $y^u = \varphi$ on ∂D and $y^u \geqq \varphi$ on Ω_0.

* Received by the editors May 8, 1989; accepted for publication (in revised form) November 30, 1990.

† Department of Mathematics, University of Iaşi, 6600 Iaşi, Romania. The author is currently at the Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio 45221-0025.

‡ Institute for Mathematics, Academy 14, Bucharest, Romania.

Here $y \in H^m(\Omega_0)$ is the solution to the elliptic boundary value problem

$$\Delta y = 0 \quad \text{in}\, \Omega_0, \tag{1.4}$$
$$\frac{\partial y}{\partial \nu} = \frac{\partial \varphi}{\partial \nu} \quad \text{in}\, \partial D, \quad y = u \quad \text{in}\, \partial \Omega,$$

where $\partial/\partial\nu$ is the inward normal derivative to ∂D.

We note that under our assumptions by standard existence results on elliptic boundary value problems, for every $u \in H^{m-1/2}(\partial\Omega)$, problem (1.4) has a unique solution $y^u \in H^m(\Omega_0) \subset C(\bar{\Omega}_0)$ (see [3]).

The relationship between problems (1.3) and (P) is clarified by Lemma 1.1 below.

LEMMA 1.1. *Assume that* $\Delta\varphi \leqq 0$ *in* D. *If* $y \in K_u$ *is the solution to problem* (1.4), *then the function* $\tilde{y}$ *defined by*

$$\tilde{y}(x) = y(x) \quad \text{if}\, x \in \Omega_0, \qquad y(x) = \varphi(x) \quad \text{if}\, x \in \bar{D}, \tag{1.5}$$

is the solution to the elliptic variational inequality (1.2).

Proof. Since ∂D is smooth we see by (1.5) that $\tilde{y} \in H^1(\Omega)$ and

$$\frac{\partial \tilde{y}}{\partial x_i} = \frac{\partial y}{\partial x_i} \quad \text{a.e. in}\, \Omega_0, \qquad \frac{\partial \tilde{y}}{\partial x_i} = \frac{\partial \varphi}{\partial x_i} \quad \text{a.e. in}\, D, \quad i = 1, 2, \cdots, n.$$

This yields, via Green's formula,

$$\begin{aligned}
\int_\Omega \nabla\tilde{y} \cdot \nabla(\tilde{y} - z)\, dx &= \int_{\Omega_0} \nabla y \cdot \nabla(y - z)\, dx + \int_D \nabla\varphi \cdot \nabla(\varphi - z)\, dx \\
&= -\int_{\Omega_0} \Delta y (y - z)\, dx - \int_{\partial D} \frac{\partial y}{\partial \nu}(y - z)\, d\sigma \\
&\quad - \int_D \Delta\varphi(\varphi - z)\, dx - \int_{\partial D} \frac{\partial \varphi}{\partial \nu}(\varphi - z)\, d\sigma \\
&= -\int_D (\varphi - z)\Delta\varphi\, dx \leqq 0 \quad \forall z \in K_u,
\end{aligned}$$

as claimed.

Lemma 1.1 shows that *problems* (1.3) and (P) *are equivalent.* However, problem (P) seems to be more convenient since it is governed by a linear elliptic equation, and so a least squares type approach will transform it into a convex control problem. Here we will use this approach to show that under suitable assumptions, problem (1.3) is approximately solvable in a sense to be explained later (see Theorems 2.1 and 2.2 below). This is quite an unexpected result if we take into account the complicated geometry of the coincidence set and it may be viewed as a controllability result for the coincidence set of the obstacle problem. In § 5 we will obtain similar results for the obstacle problem with Neumann boundary conditions.

We will use the usual notation for the spaces of continuously differentiable functions and for Sobolev spaces on Ω_0, Ω, and $\partial\Omega$.

2. The main results. Throughout this section we will assume that $\varphi \in C^m(\bar{\Omega})$ satisfies the conditions

$$\Delta\varphi(x) \leqq 0 \quad \forall x \in D, \tag{2.1}$$

$$\frac{\partial \varphi}{\partial \nu}(x) \geqq 0 \quad \forall x \in \partial D, \tag{2.2}$$

where $\partial/\partial\nu$ is the inward normal to ∂D. Assume further that ∂D is of class C^m; $m > N/2$.

THEOREM 2.1. *There is a sequence* $\{(u_n, y_n)\} \subset H^{m-1/2}(\partial\Omega) \times H^m(\Omega_0)$ *satisfying system* (1.4) *along with the constraints* $y_n \geqq \varphi$ *in* Ω_0 *and having the following property*:

(a) *For every smooth part* Γ *of* ∂D *there is no open domain* $\Pi \subset \Omega_0$ *such that* $\partial\Pi \cap \partial D = \Gamma$, $\partial\Pi \cap \partial\Omega \neq \emptyset$ *is a smooth submanifold of* $\partial\Omega$ *and*

$$y_{n_k}(x) > \varphi(x) \quad \forall x \in \Pi \tag{2.3}$$

for some subsequence $n_k \to \infty$.

Denote $\Omega_n = \{x \in \Omega_0;\ y_n(x) = \varphi(x)\}$. Theorem 2.1 shows that for every subsequence $\{n_k\} \to +\infty$ the set $\bigcap_{n_k} \Omega_{n_k}^c$ does not contain any open domain Π such that $\partial\Pi \cap \partial D$ and $\partial\Pi \cap \partial\Omega \neq \emptyset$ are smooth parts of ∂D and $\partial\Omega$, respectively (see Fig. 1).

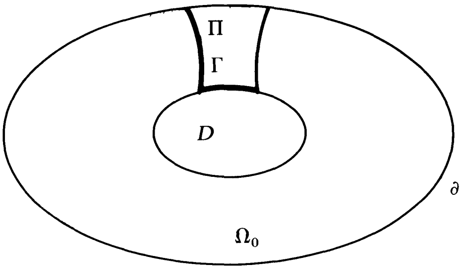


FIG. 1

Roughly speaking this means that Ω_n "asymptotically cover" the boundary ∂D. If $\Delta\varphi \leqq 0$ in D_0 then as an easy consequence of the Hopf maximum principle it follows that any open domain Ω_0' of Ω_0 having the property that $y_n = \varphi$ in $\partial\Omega_0' \backslash \partial D$, belongs to Ω_n. This fact allows us to precisely give in this case property (a), and so to give a more precise meaning to Theorem 2.1.

If $\Delta\varphi \geqq 0$ in Ω_0 then, again by the maximum principle, it follows by (1.4) that for every $u \in H^{m-1/2}(\partial\Omega)$ either $y^u \equiv \varphi$ or $y^u > \varphi$ in $\Omega_0 \cup \partial D$ and $\{x;\ y^u(x) = \varphi(x)\} \subset \partial\Omega$. Hence if $\Delta\varphi > 0$ then problem (P) has no solution.

THEOREM 2.2. *For every* $\varepsilon > 0$ *there is a connected open subset* Q_ε *of* Ω_0 *such that* $m(\Omega_0 \backslash Q_\varepsilon) \leqq \varepsilon$ *and a sequence* $\{(u_n, y_n)\} \subset H^{m-1/2}(\partial\Omega) \times H^m(\Omega_0)$ *satisfying system* (1.4) *and*

$$y_n \geqq \varphi \quad \text{a.e. in } Q_\varepsilon, \tag{2.4}$$

$$y_n \to \varphi \quad \text{weakly in } L^2(\partial D). \tag{2.5}$$

(Here m is the Lebesgue measure in R^N.)

3. Proof of Theorem 2.1. Consider the following family of optimal control problems: (P_n) Minimize $\{\int_{\partial D} y(x)\,dx + (1/2n)\|u\|^2_{H^{m-1/2}(\partial\Omega)}\}$ on all $(u, y) \in H^{m-1/2}(\partial\Omega) \times H^m(\Omega_0)$ subject to system (1.4) and to the state constraint $y(x) \geqq \varphi(x)$ for all $x \in \Omega_0$.

LEMMA 3.1. *Problem* (P_n) *has unique solution* (u_n, y_n). *Moreover, there are the functions* $p_n \in W^{1,q}(\Omega_0)$, $1 \leqq q < N/(N-1)$ *and the Borelian measures* $\mu_n \in M(\bar{\Omega}_0)$ *that satisfy the system*

$$\begin{gathered} \Delta y_n = 0 \quad \text{in } \Omega_0, \\ \frac{\partial y_n}{\partial\nu} = \frac{\partial\varphi}{\partial\nu} \quad \text{in } \partial D, \quad y_n = u_n \quad \text{in } \partial\Omega, \end{gathered} \tag{3.1}$$

$$-\Delta p_n = \mu_n \quad in\ \Omega_0,$$

$$\frac{\partial p_n}{\partial \nu} = 1 \quad in\ \partial D, \quad p_n = 0 \quad in\ \partial\Omega, \tag{3.2}$$

$$\frac{\partial p_n}{\partial \nu} = n^{-1} F u_n \quad in\ \partial\Omega, \tag{3.3}$$

$$\mu_n(y_n - z) \geqq 0 \quad \forall z \in C(\bar{\Omega}_0), \qquad z \geqq \varphi \quad on\ \bar{\Omega}_0. \tag{3.4}$$

Here $F: H^{m-1/2}(\partial\Omega) \to H^{-m+1/2}(\partial\Omega)$ is the canonical isomorphism of the space $H^{m-1/2}(\partial\Omega)$ onto its dual, and (3.2) is considered in the following weak sense:

$$\mu_n(\psi) - \int_{\Omega_0} \nabla p_n \cdot \nabla \psi \, dx + \int_{\partial D} \psi \, d\sigma + n^{-1}\langle F u_n, \psi \rangle = 0 \quad \forall \psi \in H^m(\Omega_0), \tag{3.5}$$

$$p_n = 0 \quad \text{in } \partial\Omega, \tag{3.6}$$

where $\langle \cdot, \cdot \rangle$ is the duality pairing between $H^{m-1/2}(\partial\Omega)$ and $H^{-m+1/2}(\partial\Omega)$. We note that since $p_n \in W^{1,q}(\Omega_0)$ the trace of p_n on ∂D belongs to $W^{1-1/q,q}(\partial D)$, and so (3.6) makes sense.

Proof of Lemma 3.1. We will show that there is at least one admissible pair (u, y). Then by standard division it follows that problem (P_n) has a unique solution (u_n, y_n). Moreover, we have for $\lambda \to 0$

$$u_\lambda \to u_n \quad \text{strongly in } H^{m-1/2}(\partial\Omega), \tag{3.7}$$

$$y_\lambda \to y_n \quad \text{strongly in } H^m(\Omega_0) \subset C(\bar{\Omega}_0), \tag{3.8}$$

where (u_λ, y_λ) is the solution to the following penalized problem:

(3.9) Minimize $\{\int_{\partial D} y \, d\sigma + (1/2n)\|u\|^2_{H^{m-1/2}(\partial\Omega)} + (1/2\lambda)\int_{\Omega_0} |(y-\varphi)^-|^2 \, dx\}$ subject to (1.4).

It is readily seen that there exists $p_\lambda \in H^2(\Omega_0)$ such that

$$-\Delta p_\lambda = \beta_\lambda(y_\lambda - \varphi) \quad \text{in } \Omega_0,$$

$$\frac{\partial p_\lambda}{\partial \nu} = 1 \quad \text{in } \partial D, \quad p_\lambda = 0 \quad \text{in } \partial\Omega, \tag{3.10}$$

$$\frac{\partial p_\lambda}{\partial \nu} = n^{-1} F u_\lambda \quad \text{on } \partial\Omega, \tag{3.11}$$

where $\beta_\lambda(r) = -\lambda^{-1} r^-$, for all $r \in R$.

Now let us remark for later use that there are $u_0 \in H^{m-1/2}(\partial\Omega)$ and $y_0 \in H^m(\Omega_0)$ such that $y_0(x) > \varphi(x)$, for all $x \in \bar{\Omega}_0$ and

$$\Delta y_0 = 0 \quad \text{in } \Omega_0,$$

$$\frac{\partial y_0}{\partial \nu} = \frac{\partial \varphi}{\partial \nu} \quad \text{in } \partial D, \quad y_0 = u_0 \quad \text{in } \partial\Omega. \tag{3.12}$$

Indeed, it suffices to take $u_0 \in H^{m-1/2}(\partial\Omega) \cap C^1(\partial\Omega)$ such that

$$\inf_{\partial\Omega} u_0 > \sup_{\bar{\Omega}_0} \varphi.$$

If y_0 is the corresponding solution to problem (3.12), then it follows by assumption (2.2) and the strong maximum principle that the infimum of y_0 on $\bar{\Omega}_0$ is attained on $\partial\Omega$ only. Hence $\underline{\inf}_{\Omega_0} y_0 > \underline{\sup}_{\Omega_0} \varphi$, as claimed.

Thus there is $\rho>0$ such that $y_0+\rho w \geqq \varphi$ on Ω_0 for all $w\in L^\infty(\Omega_0)$ with $\|w\|_{L^\infty(\Omega_0)}\leqq 1$. Multiplying (3.10) by $y-y_0-\rho w$ and integrating on Ω_0, we get, after some calculation involving Green's formula,

$$\begin{aligned}\rho\int_{\Omega_0}\Delta p_\lambda w\,dx &\leqq -\int_{\Omega_0}(y_\lambda-y_0)\Delta p_\lambda\,dx\\ &=-\int_{\partial D}(y_\lambda-y_0)\,d\sigma-\int_{\partial\Omega}(u_\lambda-u_0)\frac{\partial p_\lambda}{\partial\nu}\\ &=-\int_{\partial D}(y_\lambda-y_0)\,d\sigma-n^{-1}\langle Fu_\lambda,u_\lambda-u_0\rangle\end{aligned}$$

because

$$\beta_\lambda(y_\lambda-\varphi)(y_\lambda-y_0-\rho w)\geqq\frac{1}{2\lambda}(|(y_\lambda-\varphi)^-|^2-|(y_0+\rho w-\varphi)^-|^2).$$

Hence

$$\int_{\Omega_0}|\Delta p_\lambda|\,dx\leqq C\quad\forall\lambda>0,\tag{3.13}$$

where C is a positive constant independent of λ and n.

This implies that (see [3])

$$\|p_\lambda\|_{W^{1,q}(\Omega_0)}\leqq C\quad\forall\lambda>0,\tag{3.14}$$

where $1\leqq q<N/(N-1)$. Hence there is $p_n\in W^{1,q}(\Omega_0)$ and $\mu_n\in M(\bar\Omega_0)$ (the space of all Borel measures on $\bar\Omega_0$) such that on a subsequence $\lambda_k\to 0$

$$p_{\lambda_k}\to p_n\quad\text{weakly in } w^{1,q}(\Omega_0)$$

and on a subnet (generalized sequence) of $\{\beta_\lambda(y_\lambda-\varphi)\}$,

$$\beta_\lambda(y_\lambda-\varphi)\to\mu_n\quad\text{weak star in } M(\bar\Omega_0).$$

Then, letting λ tend to zero in (3.10) and (3.11), we see that y_n, u_n, p_n, and μ_n satisfy (3.1)-(3.4), thereby completing the proof of Lemma 3.1. Moreover, by estimates (3.13) and (3.14) we see that

$$\|\mu_n\|_{M(\bar\Omega_0)}+\|p_n\|_{W^{1,q}(\Omega_0)}\leqq C\quad\forall n.\tag{3.15}$$

Let us show that (u_n, y_n) satisfy the conditions of Theorem 2.1.

Let us assume that there is a domain $\Pi\subset\Omega_0$ satisfying condition (a) and argue from this to a contradiction. Since as readily seen by (3.4) support $\mu_n\subset\{x;\ y_n(x)=\varphi(x)\}$ we have

$$\begin{gathered}-\Delta p_{n_k}=0\quad\text{in }\Pi,\\ \frac{\partial p_{n_k}}{\partial\nu}=1\quad\text{in }\partial\Pi\cap\partial D,\quad p_{n_k}=0\quad\text{in }\partial\Pi\cap\partial\Omega,\end{gathered}\tag{3.16}$$

and

$$\frac{\partial p_{n_k}}{\partial\nu}=n_k^{-1}Fu_{n_k}\quad\text{in }\partial\Pi\cap\partial\Omega.\tag{3.17}$$

By estimate (3.15) we may assume without any loss of generality that

$$p_{n_k} \to p \quad \text{weakly in } W^{1,q}(\Omega_0). \tag{3.18}$$

Moreover, it is easily seen that $n_k^{-1} F u_{n_k} \to 0$ strongly in $H^{-m+1/2}(\partial\Omega)$. Thus letting n_k tend to $+\infty$ in (3.16) and (3.17), we see that p satisfies the equations

$$\begin{gathered} \Delta p = 0 \quad \text{in } \Pi, \\ \frac{\partial p}{\partial \nu} = 1 \quad \text{in } \partial\Pi \cap \partial D, \quad p = 0 \quad \text{in } \partial\Pi \cap \partial\Omega, \end{gathered} \tag{3.19}$$

$$\frac{\partial p}{\partial \nu} = 0 \quad \text{in } \partial\Pi \cap \partial\Omega. \tag{3.20}$$

Let $x_0 \in \partial\Pi \cap \partial\Omega$ be fixed, and let $r > 0$ be a sufficiently small number such that $B_{2r}(x_0) \cap \Omega \subset B_{2r}(x_0) \cap \Pi$.

Let $\alpha \in C_0^\infty(B_{2r}(x_0))$ be such that $\alpha = 1$ in $B_r(x_0)$. Then the function $\tilde{p} = \alpha p$ satisfies the equation

$$\Delta \tilde{p} = \nabla \alpha \cdot \nabla p + p \Delta \alpha \overset{\text{def}}{=} g \quad \text{in } B_{2r}(x_0) \cap \Omega = U$$

and

$$\tilde{p} = 0 \quad \text{in } \partial U.$$

Since $g \in L^q(U)$ we infer by the classical result of Agmon-Douglis-Nirenberg that $\tilde{p} \in W^{2,q}(U)$ and therefore $\nabla \tilde{p} \in L^{q^*}(U)$ where $1/q^* = 1/q - 1/N$. Repeating this argument several times and keeping in mind that $\tilde{p} = p$ on $B_r(x_0)$, it follows by standard interior and boundary regularity for elliptic equations (see, for instance, [3]) that $\tilde{p} \in C^2(\overline{B_\delta(x_0) \cap \Omega})$ where δ is suitable chosen. Since $\Delta p = 0$ in $B_\delta(x_0) \cap \Omega$ and by (3.20) $p = \partial p/\partial \nu = 0$ on $\partial\Omega \cap B_\delta(x_0)$ we conclude by the unique continuation theorem that $p = 0$ in $B_\delta(x_0) \cap \Omega$. Inasmuch as p is analytic in Π we infer that $p = 0$ in Π, and this contradicts the boundary condition $\partial p/\partial \nu = 1$ in $\partial\Pi \cap \partial D$. The contradiction we arrived at completes the proof.

Remark 3.1. To obtain an approximate solution to problem (P) we may use instead of (P_n) an optimal control of the following form:

(3.21) Minimize $\{\int_{\partial D} y(x)\, d\sigma + (1/2n)\|u\|^2_{L^2(\partial\Omega)}\}$ subject to (1.4) and state constraints $y \geqq \varphi$ in Ω_0.

Note that for every $u \in L^2(\partial\Omega)$ problem (1.4) has a unique solution $y \in L^2(\Omega_0)$ and $y \in H^2(\tilde{\Omega} \backslash \bar{D})$ for any domain $\tilde{\Omega}$, $D \subset \tilde{\Omega} \subset \Omega$ so that (3.21) makes sense. The solution (u_n, y_n) to problem (3.2) can be obtained as a limit for $\lambda \to 0$ of the solution (u_λ, y_λ) to the approximating optimality system (3.10) where $u_\lambda = n(\partial p_\lambda/\partial \nu)$.

4. Proof of Theorem 2.2. Let Q_ε be a connected open subset of Ω_0 such that $m((\Omega_0 \backslash Q_\varepsilon) \leqq \varepsilon$ and for a sufficiently small $\delta > 0$, $\{x \in \Omega_0;\ \text{dist}(x, \partial D) < \delta\} \cup \{x \in \Omega_0;\ \text{dist}(x, \partial\Omega) < \delta\} \subset Q_\varepsilon$.

Consider the following optimal control problem:

(4.1) Minimize $\{\frac{1}{2}\int_{\partial D} (y - \varphi)^2\, d\sigma + (1/2n)\|u\|^2_{H^{m-1/2}(\partial\Omega)}\}$ on all
$(u, y) \in H^{m-1/2}(d\Omega) \times H^m(\Omega_0)$ subject to system (1.4)
and to the state constraint $y \geqq \varphi$ in Q_ε.

Approximate problem (4.1) by the following family of penalized problems:

(4.2) Minimize $\{\frac{1}{2}\int_{\partial D} (y - \varphi)^2\, d\sigma + (1/2n)\|u\|^2_{H^{m-1/2}(\partial\Omega)} + (1/2\lambda)\int_{Q_\varepsilon} ((y - \varphi)^-)^2\, dx\}$ subject to (4.1).

If (u_λ, y_λ) is the optimal pair of problem (4.2), then we have as in the proof of Lemma 3.1 that

$$\begin{aligned} &u_\lambda \to u_n \quad \text{strongly in } H^{m-1/2}(\partial\Omega), \\ &y_\lambda \to y_n \quad \text{strongly in } H^m(\Omega_0), \end{aligned} \tag{4.3}$$

and

$$\begin{aligned} &-\Delta p_\lambda = \chi_\varepsilon \beta_\lambda(y_\lambda - \varphi) \quad \text{in } \Omega_0, \\ &\frac{\partial p_\lambda}{\partial \nu} = y_\lambda - \varphi \quad \text{in } \partial D, \quad p_\lambda = 0 \quad \text{in } \partial\Omega, \\ &\frac{\partial p_\lambda}{\partial \nu} = n^{-1} F u_\lambda \quad \text{in } \partial\Omega, \end{aligned} \tag{4.4}$$

where (u_n, y_n) is the solution to problem (4.1), $p_\lambda \in H^2(\Omega_0)$, and χ_ε is the characteristic function of the domain Q_ε.

Then, arguing as in the proof of Theorem 2.1, we conclude that $\{p_\lambda\}$ is bounded in $W^{1,q}(\Omega_0)$, $\{\Delta p_\lambda\}$ is bounded in $L^1(\Omega_0)$; therefore there are $p_n \in W^{1,q}(\Omega_0)$ and $\mu_n \in M(\bar{\Omega}_0)$ such that

$$\begin{aligned} &-\Delta p_n = \chi_\varepsilon \mu_n \quad \text{in } \Omega_0, \\ &\frac{\partial p_n}{\partial \nu} = y_n - \varphi \quad \text{in } \partial D, \quad y_n = 0 \quad \text{in } \partial\Omega, \\ &\frac{\partial p_n}{\partial \nu} = n^{-1} F u_n \quad \text{in } \partial\Omega. \end{aligned} \tag{4.5}$$

For $n \to \infty$, $\{y_n\}$ is bounded in $L^2(\partial D)$ and by estimate (3.15), $\{\mu_n\}$ is bounded in $M(\bar{\Omega}_0\}$, and $\{p_n\}$ is bounded in $W^{1,q}(\Omega_0)$. Thus there are $p \in W^{1,q}(\Omega_0)$ and $\mu \in M(\bar{\Omega}_0)$ such that

$$\begin{aligned} &-\Delta p = \chi_\varepsilon \mu \quad \text{in } \Omega_0, \\ &\frac{\partial p}{\partial \nu} = g \quad \text{in } \partial D, \quad p = 0 \quad \text{in } \partial\Omega, \\ &\frac{\partial p}{\partial \nu} = 0 \quad \text{in } \partial\Omega, \end{aligned} \tag{4.6}$$

where $g = w - \lim y_{n_k} - \varphi$ in $L^2(\partial D)$. In other words,

$$(\mu\chi_\varepsilon)(\psi) = \int_{\Omega_0} \nabla p \cdot \nabla \psi \, dx - \int_{\partial D} \psi g \, d\sigma \quad \forall \psi \in H^m(\Omega_0). \tag{4.7}$$

Hence

$$\begin{aligned} &-\Delta p = 0 \quad \text{in } \Omega_0 \backslash \bar{Q}\varepsilon, \\ &p = 0, \quad \frac{\partial p}{\partial \nu} = 0 \quad \text{in } \partial\Omega. \end{aligned}$$

Then, arguing as in the proof of Theorem 2.1, we deduce by the unique continuation theorem that $p = 0$ in $\Omega_0 \backslash \bar{Q}_\varepsilon$. Since p is analytic in $\Omega_0 \backslash \bar{Q}_\varepsilon$ and consequently in a neighborhood of ∂D we infer that $p \equiv 0$ in v. If in (4.7) we take supp $\psi \subset \Omega_0 \backslash \bar{Q}_\varepsilon$ we see that $g \equiv 0$, as claimed.

Clearly this argument implies that $y_n \to \varphi$ weakly in $L^2(\partial D)$, thereby completing the proof.

5. Neumann boundary conditions. Consider the controllability problem (1.3) for the variational inequality

$$\begin{aligned} &y-\Delta y=0 \quad \text{in } \{x\in\Omega;\, y(x)>\varphi(x)\},\\ &y-\Delta y\leqq 0 \quad \text{in } \Omega, \quad y\geqq\varphi \quad \text{in } \Omega, \\ &\frac{\partial}{\partial \nu}y=u \quad \text{in } \partial\Omega, \end{aligned} \tag{5.1}$$

where $u\in H^{m-3/2}(\partial\Omega)$, i.e.,

(5.2) Find $u\in H^{m-3/2}(\partial\Omega)$ such that $D\subset E_u$, where D is a given smooth subset of Ω and E_u is the coincidence set associated with problem (4.1).

We will assume as above that $\varphi\in C^m(\bar{\Omega})$, ∂D is of class C^m, and

$$\begin{aligned} &\varphi-\Delta\varphi\geqq 0 \quad \text{in } D,\\ &\frac{\partial\varphi}{\partial\nu}\geqq 0 \quad \text{in } \partial D. \end{aligned} \tag{5.3}$$

Then, arguing as in the proof of Lemma 1.1, which remains true in this case, we may reduce problem (4.2) to

(P_1) Find $u\in H^{m-3/2}(\partial\Omega)$ such that $y^u=\varphi$ on ∂D and $y^u\geqq\varphi$ on Ω, where y^u is the solution to boundary value problem

$$\begin{aligned} &y-\Delta y=0 \quad \text{in } \Omega_0,\\ &\frac{\partial y}{\partial\nu}=\frac{\partial\varphi}{\partial\nu} \quad \text{in } \partial D, \qquad \frac{\partial y}{\partial\nu}=u \quad \text{in } \partial\Omega. \end{aligned} \tag{5.4}$$

Also in this case we have a controllability result of the nature of Theorem 2.1.

THEOREM 5.1. *There is a sequence* $(u_n, y_n)\subset H^{m-3/2}(\partial\Omega)\times H^m(\Omega_0)$ *satisfying system* (5.4) *and the constraints* $y_n\geqq\varphi$ *in* Ω_0, *and having property* (a) *of Theorem* 2.1.

If it is assumed further that

$$\varphi-\Delta\varphi\geqq 0 \quad \text{in } \Omega_0, \tag{5.5}$$

then every subdomain Ω_0' of Ω_0 such that $y_n=\varphi$ on $\partial\Omega_0'\backslash\partial D$ belongs to Ω_n.

In other words, u_n is an approximating solution to controllability problem $\{u\in H^{m-3/2}(\partial\Omega);\, E_u\supset D\}$.

Since the proof of Theorem 5.1 is similar to that of Theorem 2.1, it will only be sketched.

Consider the following optimal control problem:

(P_n^1) Minimize $\{\int_{\partial D} y\, d\sigma+(1/2n)\|u\|^2_{H^{m-3/2}(\partial\Omega)}\}$ on all $(y,u)\in H^m(\Omega_0)\times H^{m-3/2}(\partial\Omega)$ subject to (5.4) and to state constraint $y\geqq\varphi$ on Ω_0.

We associate the following penalized problem:

(5.6) Minimize $\{\int_{\partial D} y\, d\sigma+(1/2n)\|u\|^2_{H^{m-3/2}(\partial\Omega)}+(1/2\lambda)\int_{\Omega_0}[(y-\varphi)]^2\,dx\}$ subject to (5.4)

and denote (u_n, y_n), (u_λ, y_λ) the corresponding solution to problem (P_n) and (5.6), respectively.

We have for $\lambda\to 0$

$$u_\lambda \to u_n \quad \text{strongly in } H^{m-3/2}(\partial\Omega), \tag{5.7}$$

$$y_\lambda \to y_n \quad \text{strongly in } H^{m}(\Omega_0) \tag{5.8}$$

and

$$\begin{gathered} p_\lambda - \Delta p_\lambda = \beta_\lambda(y_\lambda - \varphi) \quad \text{in } \Omega_0, \\ \frac{\partial p_\lambda}{\partial \nu} = 1 \quad \text{in } \partial D, \quad \frac{\partial p_\lambda}{\partial \nu} = 0 \quad \text{in } \partial\Omega, \end{gathered} \tag{5.9}$$

$$p_\lambda + n^{-1}\phi(u_\lambda) = 0 \quad \text{in } \partial\Omega, \tag{5.10}$$

where $\phi: H^{m-3/2}(\partial\Omega)\to H^{3/2-m}(\partial\Omega)$ is the canonical isomorphism of the space $H^{m-3/2}(\partial\Omega)$ onto its dual $H^{3/2-m}(\partial\Omega)$.

If (u_0, y_0) is the pair defined as in the proof of Theorem 2.1, then $\tilde{u}_0 = \partial y_0/\partial \nu$ is an admissible control for problem (P_n) and we find, as there, the estimate

$$\begin{aligned} \rho \int_{\Omega_0} |p_\lambda - \Delta p_\lambda|\, dx &\leqq -\int_{\partial D} (y_\lambda - y_0)\, d\sigma + \int_{\partial D} p_\lambda \frac{\partial}{\partial \nu}(y_\lambda - y_0)\, d\sigma \\ &= -\int_{\partial D} (y_\lambda - y_0)\, d\sigma - n^{-1} < \phi(u_\lambda), u_\lambda - u_0 > \leqq C, \end{aligned} \tag{5.11}$$

where C is independent of λ and n. (Here $\langle\cdot,\cdot\rangle$ is the duality pairing between $H^{m-3/2}(\partial\Omega)$ and $H^{3/2-m}(\partial\Omega)$.) Hence $\{p_\lambda - \Delta p_\lambda\}$ is bounded in $L^1(\Omega_0)$ and so by Lemma 2.3 of [3] it follows that $\{p_\lambda\}$ is bounded in $W^{1,q}(\Omega_0)$, where $1\leqq q< N(N-1)$. Thus we may pass to limit in (5.9) to get the optimality system associated with (P_n)

$$\begin{gathered} y_n - \Delta y_n = 0 \quad \text{in } \Omega_0, \\ \frac{\partial y_n}{\partial \nu} = \frac{\partial \varphi}{\partial \nu} \quad \text{in } \partial D, \quad \frac{\partial y_n}{\partial \nu} = u_n \quad \text{in } \partial\Omega, \end{gathered} \tag{5.12}$$

$$\begin{gathered} p_n - \Delta p_n = \mu_n \quad \text{in } \Omega_0, \\ \frac{\partial p_n}{\partial \nu} = 1 \quad \text{in } \partial D, \quad \frac{\partial p_n}{\partial \nu} = 0 \quad \text{in } \partial\Omega, \end{gathered} \tag{5.13}$$

$$p_n + n^{-1}\phi(u_n) = 0 \quad \text{in } \partial\Omega, \tag{5.14}$$

$$\mu_n \in M(\bar{\Omega}_0), \quad \mu_n(y_n - z) \geqq 0, \quad \forall z \in C(\bar{\Omega}_0), \quad z \geqq \varphi. \tag{5.15}$$

The solution $p_n \in W^{1,q}(\Omega_0)$ to system (5.13) should be understood, of course, in the weak sense (3.5), i.e.,

$$\mu_n(\psi) - \int_{\Omega_0} \nabla p_n \nabla \psi\, dx - \int_{\partial D} \psi\, d\sigma = 0, \quad \forall \psi \in H^m(\Omega_0).$$

Note also that the estimate (3.15) remains valid in this case. Arguing as in the proof of Theorem 2.1, it follows that (u_n, y_n) has the property required for Ω_n.

Now, if there is $\Pi \subset \Omega_0$ satisfying condition (a) of Theorem 2.1, we have

$$p_{n_k} - \Delta p_{n_k} = 0 \quad \text{in } \Pi,$$

$$\frac{\partial p_{n_k}}{\partial \nu} = 1 \quad \text{in } \partial\Pi \cap \partial D, \quad \frac{\partial p_{n_k}}{\partial \nu} = 0 \quad \text{in } \partial\Pi \cap \partial\Omega,$$

$$p_{n_k} = n_k^{-1}\phi(u_{n_k}) \quad \text{in } \partial\Pi \cap \partial\Omega,$$

and letting $n_k \to +\infty$ we see that there is $p \in W^{1,q}(\Omega_0)$ such that

$$p - \Delta p = 0 \quad \text{in } \Pi,$$

$$\frac{\partial p}{\partial \nu} = 1 \quad \text{in } \partial\Pi \cap \partial D, \quad \frac{\partial p}{\partial \nu} = 0 \quad \text{in } \partial\Pi \cap \partial\Omega,$$

$$p = 0 \quad \text{in } \partial\Pi \cap \partial\Omega.$$

As seen in the proof of Theorem 2.1 this implies, via regularity theory for elliptic boundary value problems and the unique continuation theorem, that $p = 0$ in Π, thereby completing the proof.

The theorem below follows by a similar argument.

THEOREM 5.2. *For every $\varepsilon > 0$, there is an open connected subset Q_ε of Ω_0 such that $m(\Omega_0 \backslash Q_\varepsilon) \leqq \varepsilon$ and a sequence $\{(u_n, y_n)\} \subset H^{m-3/2}(\partial\Omega) \times H^m(\Omega_0)$ satisfying system* (5.4) *and conditions* (2.4), (2.5) *in Theorem* 2.2.

REFERENCES

[1] V. BARBU, *Optimal Control of Variational Inequalities*, Research Notes in Mathematics 100, Pitman, Boston, London, Melbourne, 1984.

[2] H. BREZIS AND W. A. STRAUSS, *Semilinear second order elliptic equations in L^1*, J. Math. Soc. Japan, 25 (1973), pp. 565-590.

[3] J. L. LIONS AND E. MAGENES, *Nonhomogeneous Boundary Value Problems and Applications*, Springer-Verlag, Berlin, Heidelberg, New York, 1972.

[4] CH. SAGUEZ, *Contrôle optimal de systèmes à frontière libre*, Thèse l'Université de Technologie de Compiègne, 1980.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 5, pp. 1160-1175, September 1991

008

A SUBSPACE DECOMPOSITION PRINCIPLE FOR SCALED GRADIENT PROJECTION METHODS: GLOBAL THEORY*

J. C. DUNN†

Abstract. Fast gradient projection methods for constrained minimization problems, $\min_\Omega J$, achieve their superior asymptotic convergence rates by scaling the objective function gradient $\nabla J(u)$ prior to the projection step. The scaling procedure investigated in this article decomposes $\nabla J(u)$ relative to specially constructed complementary orthogonal subspaces N and T, multiplies the N-component of $\nabla J(u)$ by a positive scalar $s_N(u)$, transforms the T-component of $\nabla J(u)$ with a bounded linear operator $S_T: T \to T$, and adds the resulting vectors to obtain the scaled gradient of J at u. A global convergence analysis is undertaken here for scaled gradient projection (SGP) methods that utilize this technique and a compatible steplength rule of the Bertsekas-Gafni type in closed convex Ω defined by m smooth inequality constraints. Extensive comparisons are drawn with a related SGP method based on the Gafni-Bertsekas dual cone decomposition technique. The two SGP schemes have similar global convergence properties, but the subspace decomposition procedure is generally easier to implement, and is better suited to convergence acceleration in nonpolyhedral Ω. A comprehensive local convergence theory has also been constructed for the new scheme and will be presented in a sequel to this article. In general, SGP algorithms are well suited to optimal control problems, network flow problems, and other cases where Ω is a Cartesian product of simple convex sets.

Key words. gradient projection, scaling, convergence acceleration

AMS(MOS) subject classifications. 49D07, 65K10, 65B99

1. Introduction.

The gradient projection (GP) methods in [1]-[10] and the more recent scaled gradient projection (SGP) algorithms in [7] and [11]-[15] are useful for a variety of specially structured constrained minimization problems

$$\min_{u \in \Omega} J(u),$$

where J is a smooth real function on a real Hilbert space $\{\mathcal{U}, \langle \cdot, \cdot \rangle\}$ and Ω is a nonempty closed convex set in $\mathcal{U}$. These algorithms generate successive feasible approximations $u^i \in \Omega$ recursively with

$$u^{i+1} = P_\Omega(u^i + \sigma^i v^i), \tag{1a}$$

where at each $u \in \Omega$,

$$v = v_N + v_T, \tag{1b}$$

$$v_N = s_N P_N(-\nabla J(u)), \tag{1c}$$

$$v_T = P_T S_T P_{N^*}(-\nabla J(u)), \tag{1d}$$

$$N = \text{a closed convex cone containing } K(u), \tag{1e}$$

$$K(u) = \text{the normal cone at } u = \{w: \forall v \in \Omega, \langle w, v-u \rangle \leqq 0\}, \tag{1f}$$

$$\mathrm{N}^* = \text{the dual cone for } N = \{w^*: \forall w \in N, \langle w^*, w \rangle \leqq 0\}, \tag{1g}$$

$$T = \{v_N\}^\perp \cap N^*, \tag{1h}$$

$$[T] = \text{the closed linear hull of } T, \tag{1i}$$

* Received by the editors October 24, 1988; accepted for publication (in revised form) September 11, 1990. This investigation was supported by National Science Foundation research grant DMS8702929.

† Mathematics Department, Box 8205, North Carolina State University, Raleigh, North Carolina 27695-8205.

(1j) $$\mu_3 \geqq s_N \geqq \mu_2 > 0,$$

$$S_T = \text{a bounded linear map from } [T] \text{ into } [T] \text{ such that}$$

(1k) $$\langle P_{N^*}(-\nabla J(u)), S_T P_{N^*}(-\nabla J(u))\rangle \geqq \mu_0 \|P_{N^*}(-\nabla J(u))\|^2$$

(1l) $$\|S_T\| \leqq \mu_1,$$

$$\sigma = \max s, \quad \text{subject to}$$

$$\frac{s}{\alpha} \in \{1, \beta, \beta^2, \cdots\},$$

and

(1m) $$J(u) - J(P_\Omega(u+sv)) \geqq \delta\{(s_N s)^{-1}\|u + sv_T - P_\Omega(u+sv)\|^2 + \langle P_{N^*}(-\nabla J(u)), v_T\rangle s\},$$

and where δ and β are fixed real numbers in (0, 1), α and $\mu_0, \cdots, \mu_3$ are fixed positive real numbers, and P_A denotes projection into the set A (relative to the fixed metric in $\mathcal{U}$). In practice, N^i, s_N^i, and S_T^i are further restricted at each iteration by auxiliary rules designed to control global and local convergence behavior.

When $N = \mathcal{U}$ and $s_N = 1$, the scheme (1) reduces to a simple Goldstein-Levitin-Polyak iteration [1], [2],

$$u^{i+1} = P_\Omega(u^i - \sigma^i \nabla J(u^i))$$

with a steplength rule of the Bertsekas type [3]. In this case, the corresponding iterate sequences $\{u^i\}$ behave globally and locally like their steepest descent counterparts for unconstrained minimization [16], [17]. More specifically, $\{J(u^i)\}$ is generally decreasing, limit points of $\{u^i\}$ are stationary, and u^i typically becomes "more stationary" with increasing i; however, the asymptotic rates of improvement may be quite poor, depending on the local structure of J and Ω [5], [6]. On the other hand, if $\Omega = \mathcal{U}$ and $N = \{0\} = K(u)$, then (1) reduces to an unconstrained variable metric gradient iteration,

$$u^{i+1} = u^i - \sigma^i S^i \nabla J(u^i)$$

with steplengths σ of the Armijo-Goldstein type [18]-[20]. Under these circumstances, it is possible to secure both the desired global convergence properties *and* fast asymptotic convergence rates by fixing δ in $(0, \frac{1}{2})$, setting $\alpha = 1$, and making sure that the scaling operators S^i approximate the Newtonian transformations $(\nabla^2 J(u^i))^{-1}$ on the span of $\{\nabla J(u^i)\}$ when u^i is near a nonsingular local minimizer of J [16], [17], [20]. Hence for general closed convex $\Omega \subset \mathcal{U}$, the question is this: Are there cones N large enough to preserve rudimentary descent and limit point stationarity properties for (1), yet small enough to ensure that the corresponding scaling subspaces $[T]$ will support Newton-like convergence-accelerating operators S_T? Bertsekas [11] and Gafni and Bertsekas [12] were the first to show that these two antagonistic requirements can indeed be satisfied in polyhedral convex sets Ω. The present article analyzes an alternative to the construction in [12] outlined previously in [15]; this alternative SGP scheme works in polyhedral Ω and also in nonpolyhedral Ω prescribed by finitely many smooth nonaffine inequality constraints.

The general cone construction in [12] proceeds from an affine inequality representation

(2a) $$\Omega = \{u : \forall j \in \mathcal{J}, \langle a^j, u\rangle - b^j \leqq 0\},$$

where $\mathcal{J}$ is an index set, $a^j \in \mathcal{U}$, and $b^j \in \mathbb{R}$; in a Hilbert space, every closed convex Ω has such a representation, with $\mathcal{J}$ finite only if Ω is polyhedral. Let ε_0 be an arbitrary but fixed positive real number, and at each $u \in \Omega$, put

(2b) $$d(u) = \|u - P_\Omega(u - \nabla J(u))\|,$$

(2c) $$\varepsilon(u) = \min\{\varepsilon_0, d(u)\},$$

$$\mathcal{J}(u)=\{j\in\mathcal{J}: \langle a^j, u\rangle - b^j \geqq -\|a^j\|\,\varepsilon(u)\}, \tag{2d}$$

$$C(u)=\{w: \forall j\in\mathcal{J}(u), \langle w, a^j\rangle \leqq 0\}. \tag{2e}$$

If $d(u)=0$, then u is *stationary*, i.e.,

$$\forall v\in\Omega, \quad \langle \nabla J(u), v-u\rangle \geqq 0,$$

or equivalently,

$$-\nabla J(u)\in K(u).$$

In this case, any closed convex cone $N\supset K(u)$ will serve for (1). If $d(u)\neq 0$, then u is not stationary and [12] employs cones N satisfying

$$N\supset C(u)^*=\{w^*: \forall w\in C(u), \langle w^*, w\rangle\leqq 0\} \tag{2f}$$

(actually, [12] imposes the somewhat stronger condition (13) in § 3).

The cone $C(u)^*$ is expressly designed to contain not only $K(u)$ but all neighboring normal cones $K(u')$ for $u'\in\Omega$ near u, and it is this feature that ultimately yields the robust global limit point stationarity theorem for (1)-(2) in arbitrary closed convex Ω [12]. Furthermore, if Ω is polyhedral and u is sufficiently close to a nonsingular local minimizer $\bar u$ (i.e., a $\bar u$ satisfying standard Kuhn-Tucker second-order sufficient conditions), then $C(u)^*=K(\bar u)$, and the cone T corresponding to the lower bound $N=C(u)^*$ in (2f) is just the tangent space $K(\bar u)^\perp$ for the unique polyhedral face $\mathcal{F}$ containing $\bar u$ in its relative interior ri $\mathcal{F}$. For $N^i\supset C(u^i)^*$, any sequence $\{u^i\}$ generated by (1)-(2) will converge to $\bar u$ from nearby starting points and eventually enter and remain within ri $\mathcal{F}$. In effect, these processes terminate in an "unconstrained" variable metric gradient iteration for J restricted to a translate of $K(\bar u)^\perp = T$. Hence in polyhedral Ω, (1)-(2) is locally linearly convergent to nonsingular minimizers $\bar u$, and can be *superlinearly* convergent if $N^i=C(u^i)^*$, $\delta\in(0,\frac12)$, $\alpha=1$, and the scaling operators S^i_T asymptotically approximate the inverse reduced Hessians

$$P_{T^i}\nabla^2 J(u^i)|_{T^i}^{-1}$$

in the one-dimensional subspaces generated by $P_{T^i}\nabla J(u^i)$. On the other hand, no comparable convergence acceleration claim is made in [12] for nonpolyhedral Ω, and it is unlikely that such a result can be proved. For example, suppose that $\mathcal{U}=\mathbb{R}^3$ and Ω is the unit ball $B(0,1)=\{u: \|u\|\leqq 1\}$ with affine inequality representation

$$B(0,1)=\{u: \forall j\in\mathcal{J}, \langle j, u\rangle - 1\leqq 0\},$$

where

$$\mathcal{J}=S(0,1)=\{j\in\mathcal{U}: \|j\|=1\}.$$

Near any nonsingular boundary point minimizer $\bar u\in S(0,1)$, it seems clear that Newton-like convergence acceleration is possible only if $[T]$ approximates the two-dimensional space tangent to the smooth manifold $S(0,1)$ at $\bar u$; however, for any u near but not equal to $\bar u$, the sets $C(u)$ and $C(u)^*$ in (2) are dual nondegenerate right circular cones, the cone T corresponding to $N=C(u)^*$ is typically a generator in $C(u)$, and $[T]$ is then a subspace of dimension one.

The alternative SGP formulation in [15] restricts the cone N to the family of closed *subspaces* containing $K(u)$; under these circumstances, $[T]=T=N^*=N^\perp$, and (1) reduces to the simpler subspace decomposition and scaling scheme

$$u\to P_\Omega(u+\sigma v) \tag{3a}$$

with

(3b) $$v = v_N + v_T,$$

(3c) $$v_N = -s_N P_N \nabla J(u),$$

(3d) $$v_T = -S_T P_T \nabla J(u),$$

(3e) $$N = \text{a closed subspace} \supset K(u),$$

(3f) $$K(u) = \{w\colon \forall v \in \Omega, \langle w, v-u\rangle \leqq 0\},$$

(3g) $$T = N^{\perp},$$

(3h) $$\mu_3 \geqq s_N \geqq \mu_2,$$

$$S_T = \text{a bounded linear map from } T \text{ into } T\text{, such that}$$

(3i) $$\langle P_T \nabla J(u), S_T P_T \nabla J(u)\rangle \geqq \mu_0 \|P_T \nabla J(u)\|^2$$

$$\|S_T\| \leqq \mu_1,$$

(3j) $$\sigma = \max s, \quad \text{subject to}$$

$$\frac{s}{\alpha} \in \{1, \beta, \beta^2, \cdots\}$$

and

(3k) $$J(u) - J(P_\Omega(u+sv)) \geqq \delta\{(s_N s)^{-1}\|u + sv_T - P_\Omega(u+sv)\|^2 - \langle P_T \nabla J(u), v_T\rangle s\}.$$

The subspaces N are further restricted as follows in closed convex sets with smooth inequality representations:

(4a) $$\Omega = \{u\colon g_i(u) \leqq 0, i = 1, \cdots, m\}.$$

Fix $\varepsilon_0 > 0$ and $\theta > 1$ (see Note 2 at the end of § 3) and at each $u \in \Omega$, put

(4b) $$d(u) = \|u - P_\Omega(u - \nabla J(u))\|,$$

(4c) $$\varepsilon(u) = \min\{\varepsilon_0, d(u)\},$$

(4d) $$\mathcal{J}_0(u) = \{j\colon g_j(u) = 0\},$$

(4e) $$\mathcal{J}(u) = \{j\colon g_j(u) \geqq -\theta\|\nabla g_j(u)\|\varepsilon(u)\} \supset \mathcal{J}_0(u),$$

(4f) $$\mathcal{G}_0(u) = \{\nabla g_j(u)\}_{j \in \mathcal{J}_0(u)},$$

(4g) $$\mathcal{G}(u) = \{\nabla g_j(u)\}_{j \in \mathcal{J}(u)} \supset \mathcal{G}_0(u),$$

(4h) $$N_0(u) = [\mathcal{G}_0(u) \cup \{0\}],$$

(4i) $$N(u) = [\mathcal{G}(u) \cup \{0\}] \supset N_0(u).$$

(Observe that the sets $\mathcal{J}_0(u)$, $\mathcal{J}(u)$, $\mathcal{G}_0(u)$, and $\mathcal{G}(u)$ may be empty at certain $u \in \Omega$.) Then [15] requires that

(4j) $$N \supset N(u).$$

This condition implies (3e) under the standard assumptions set forth in the following.

Constraint Qualification (Q). Either

(i) The functions g_j in (4a) are affine, i.e., $g_j(u) = \langle a^j, u\rangle - b^j$, with $a^j \in \mathcal{U}$ and $b^j \in \mathbb{R}$; or

(ii) Equation (4a) is a *normal representation* for Ω, i.e., for all $u \in \Omega$, $\mathcal{G}_0(u)$ is linearly independent.

Unless otherwise noted, it will be assumed that (4a) satisfies (Q).

When (Q) holds, the Hilbert space extension of the Farkas lemma [21] guarantees that $K(u)$ consists of zero and all nonnegative linear combinations of vectors in $\mathcal{G}_0(u)$, and hence that $N(u) \supset K(u)$. Moreover, the normality condition (Q)(ii) typically does hold when Ω is the kind of convex set for which the projections in (3) are readily implemented, e.g., Cartesian products of orthants, boxes, simplices, balls, cylinders, convex figures of revolution, and the like (feasible sets of this type are commonplace in optimal control and network flow applications [11]-[15]). On the other hand, in nonpolyhedral sets (4a), the subspace $N(u)$ is not large enough to contain neighboring normal cones $K(u')$, and hence the global convergence proof strategy developed in [12] for (1)-(2) cannot be applied to (3)-(4). A different approach based on special properties of the normal cone $K(u)$ in sets (4a) is fully developed in §§ 2-4 below. The resulting limit point stationarity theorem is somewhat narrower than its counterpart in [12] for arbitrary closed convex sets; however, the scheme (3)-(4) is generally easier to implement than (1)-(2), and its local convergence behavior appears equal or superior to the behavior of (1)-(2) in closed convex sets (4a) satisfying (Q). When Ω is polyhedral, and when $N = C(u)^*$ in (1)-(2) and $N = N(u)$ in (3)-(4), the corresponding SGP iterations turn out to be locally equivalent near nondegenerate stationary points, i.e., the two schemes generate identical iterates near $\bar{u}$ even though $N(u) \neq C(u)^*$ in general. With reference to [12], this means that (3)-(4) can be locally superlinearly convergent to nonsingular minimizers $\bar{u}$ in polyhedral sets (4a) if s_N and S_T are properly chosen. More generally, in any closed convex set (4a) satisfying (Q), the iteration (3)-(4) is at least locally linearly convergent to nonsingular local minimizers $\bar{u}$, and will converge superlinearly to such $\bar{u}$'s when S_T^i asymptotically approximates certain Newtonian scaling operators described in [14] on the span of $\{P_{T_i} \nabla J(u^i)\}$. Proofs for these claims are supplied in a sequel to the present article.

The cost of computing the various projections and Newtonian scaling operations is clearly an essential practical consideration that limits the application of SGP methods to problems with specially structured feasible sets and objective functions. Nevertheless, the class of likely problems is surprisingly large and includes the k-stage input-constrained discrete-time optimal control problems treated in [11] and later in [7] and [13]-[15]. For these problems, Ω is a Cartesian product of k simple convex sets defined by separable constraints, $N(u)$ and $T(u)$ are Cartesian products of k mutually orthogonal subspaces, and Newtonian scaling can be computed cheaply with dynamic programming methods or other techniques, typically in $O(k)$ flops. Under these circumstances, $\sigma(u)$ is also readily computed, and the cost per iteration for Newtonian SGP schemes is some fixed multiple of the unscaled GP cost, with a cost multiplier that depends on the complexity of the stagewise loss functions and dynamical transformations, but does *not* depend on k. For a more detailed discussion of these points, see [14], [15], [22]. For network flow applications of the SGP scheme, see [12].

2. The Bertsekas–Gafni steplength rule. Fix u in Ω and v in $\mathcal{U}$, and put

$$\phi(\sigma) = P_\Omega(u + \sigma v)$$

for $\sigma \geqq 0$. Our immediate objective is to show that near $\sigma = 0$, the quotients $\sigma^{-1}\langle \nabla J(u), u - \phi(\sigma)\rangle$ and $\sigma^{-1}(J(u) - J(\phi(\sigma)))$ are uniformly bounded below by a certain measure of nonstationarity for u when v satisfies (1). This bound points directly to the steplength rule (1m) and is crucial in the global convergence analysis for (1)-(2) in [12] and (3)-(4) in § 4.

Since P_Ω is nonexpansive,

$$\|\phi(\sigma) - u\| \leqq \|v\| \sigma \tag{5}$$

and therefore

$$J(u) - J(\phi(\sigma)) = \langle \nabla J(u), u - \phi(\sigma)\rangle + o(\sigma). \tag{6}$$

By the Hilbert space projection theorem [23],

$$u+\sigma v-\phi(\sigma)\in K(\phi(\sigma))$$

and thus

$$\sigma\langle v, \phi(\sigma)-u\rangle \geqq \|\phi(\sigma)-u\|^2.$$

Observe that by Lemma 2.2 of [24]

$$-\nabla J(u)=P_N(-\nabla J(u))+P_{N^*}(-\nabla J(u)) \tag{7a}$$

with

$$\langle P_N(-\nabla J(u)), P_{N^*}(-\nabla J(u))\rangle=0. \tag{7b}$$

Hence, if v satisfies (1), then

$$\sigma s_N\langle \nabla J(u), u-\phi(\sigma)\rangle \geqq \|u+\sigma v_T-\phi(\sigma)\|^2-\|v_T\|^2\sigma^2+\langle x, \phi(\sigma)-u\rangle\sigma \tag{8a}$$

with

$$x=(v_T+s_N P_{N^*}(-\nabla J(u)))\in T. \tag{8b}$$

This estimate is carried further in the next two lemmas.

LEMMA 1. *Let Ω be a closed convex set. Fix u in Ω and v in $\mathcal{U}$. Then for all x and σ,*

$$x\in K(u)^* \quad \textit{and} \quad \sigma\geqq 0 \Rightarrow \langle x, \phi(\sigma)-u\rangle \geqq \langle x, v\rangle\sigma+\|x\|\eta(\sigma),$$

where

$$\eta(\sigma)=o(\sigma)$$

as $\sigma\to 0_+$.

Proof. Observe that $v-P_{K(u)^*}v\in K(u)$ and therefore $\langle x, v-P_{K(u)^*}v\rangle\leqq 0$ for x in $K(u)^*$. Furthermore, by Lemmas 2.4 and 4.6 of [24],

$$\phi(\sigma)-u=(P_{K(u)^*}v)\sigma+o(\sigma)$$

and thus

$$\langle x, \phi(\sigma)-u\rangle \geqq \langle x, v\rangle\sigma+\|x\|\cdot o(\sigma). \qquad \square$$

LEMMA 2. *Assume that Ω is a closed convex set. Fix u in Ω, suppose that J is Fréchet differentiable at u, and let v satisfy* (1b)-(11). *Then for all $\sigma>0$,*

$$\langle \nabla J(u), u-\phi(\sigma)\rangle \geqq (s_N^{-1}\sigma^{-2}\|u+\sigma v_T-\phi(\sigma)\|^2+\langle P_{N^*}(-\nabla J(u)), v_T\rangle)\sigma+s_N^{-1}\|x\|\eta(\sigma),$$

where

$$x=[v_T+s_N P_{N^*}(-\nabla J(u))]\in T$$

and

$$\eta(\sigma)=o(\sigma)$$

as $\sigma\to 0_+$. In addition, for α fixed in $(0,\infty)$ *and all $\sigma\in(0,\alpha]$,*

$$s_N^{-1}\sigma^{-2}\|u+\sigma v_T-\phi(\sigma)\|^2+\langle P_{N^*}(-\nabla J(u)), v_T\rangle \geqq d_1(u)\geqq 0,$$

where

$$d_1(u)=\begin{cases}\langle P_{N_*}(-\nabla J(u)), v_T\rangle & \textit{if } P_{N^*}(-\nabla J(u))\neq 0,\\ s_N \min\{1, (s_N\alpha)^{-2}\}\, d(u)^2, & \textit{if } P_{N^*}(-\nabla J(u))=0,\end{cases}$$

$$d(u)=\|u-P_\Omega(u-\nabla J(u))\|$$

and

$$d_1(u)=0 \Leftrightarrow u \quad \textit{is stationary.}$$

Proof. By construction, $T\subset N^*\subset K(u)^*$, $x\in T$ and $\langle x, v_N\rangle=0$. Hence the first claim follows at once from (8) and Lemma 1. The remaining claims are consequences of the coercivity condition (1k) and the following key monotonicity conditions proved in [12], [9], and [8]. For all σ, τ,

$$\tau>\sigma>0 \Rightarrow \sigma^{-1}\|\delta(\sigma)-u\| \geqq \tau^{-1}\|\phi(\tau)-u\|, \tag{9a}$$

$$\tau>\sigma>0 \Rightarrow \|\phi(\tau)-u\| \geqq \|\phi(\sigma)-u\|. \tag{9b}$$

Thus if $P_{N^*}(-\nabla J(u))\neq 0$, note that $w-P_Tw\in T^*$ for all $w\in \mathcal{U}$, and therefore

$$\begin{aligned}\langle P_{N^*}(-\nabla J(u)),\, v_T\rangle &= \langle P_{N^*}(-\nabla J(u)),\, P_TS_TP_{N^*}(-\nabla J(u))\rangle\\ &\geqq \langle P_{N^*}(-\nabla J(u)),\, S_TP_{N^*}(-\nabla J(u))\rangle\\ &\geqq \mu_0\|P_{N^*}(-\nabla J(u))\|^2\\ &>0.\end{aligned} \tag{10}$$

On the other hand, if $P_{N^*}(-\nabla J(u))=0$, then $v_T=0$ and $P_N(-\nabla J(u))=-\nabla J(u)$, and conditions (9) therefore yield

$$\begin{aligned} s_N^{-1}\sigma^{-2}\|u+\sigma v_T-\phi(\sigma)\|^2 &= s_N[(s_N\sigma)^{-2}\|u-P_\Omega(u-s_N\sigma\nabla J(u))\|^2]\\ &\geqq s_N \min\{1, (s_N\alpha)^{-2}\}\, d(u)^2,\end{aligned}$$

where $d(u)=0$ if and only if u is stationary. To complete the proof, observe that if u is stationary, then $0=\|P_{K(u)^*}(-\nabla J(u))\|\geqq\|P_{N^*}(-\nabla J(u))\|$. □

The foregoing development parallels [12] with one notable departure: in [12], the cones N are required to satisfy a condition that implies (2f) (see (13) in § 3). Under these circumstances, Lemma 1 can be strengthened and the estimates in Lemma 2 are obtained with $\eta(\sigma)=0$ for $\sigma\|v\|<\varepsilon(u)$ (see (2c)). Improvements of this kind are needed later on in the global and local convergence analyses for (1)-(2) and (3)-(4); however, Lemma 2 is strong enough to establish the following basic strict descent property in the general setting of (1).

LEMMA 3. *Let the hypotheses of Lemma* 2 *hold and fix* δ *in* (0, 1). *Then for sufficiently small* σ *in* $(0, \alpha]$,

$$\begin{aligned} J(u)-J(\phi(\sigma)) &\geqq \delta[(s_N\sigma)^{-1}\|u+\sigma v_T-\phi(\sigma)\|^2+\langle P_{N^*}(-\nabla J(u)),\, v_T\rangle\sigma]\\ &\geqq \delta d_1(u)\sigma\end{aligned} \tag{11a}$$

with $d_1(u)\geqq 0$ *and*

$$d_1(u)=0 \Leftrightarrow u \quad \textit{is stationary.} \tag{11b}$$

Proof. If u is stationary, then $P_{N^*}(-\nabla J(u))=v_T=d(u)=d_1(u)=0$, and $\phi(\sigma)=u$ for all $\sigma\geqq 0$; in this case (11a) holds trivially for all $\sigma\geqq 0$. If u is not stationary, then (6) and Lemma 2 give

$$\begin{aligned} J(u)-J(\phi(\sigma)) &\geqq [s_N^{-1}\sigma^{-2}\|u+\sigma v_T-\phi(\sigma)\|^2+\langle P_{N^*}(-\nabla J(u)),\, v_T\rangle]\sigma+o(\sigma)\\ &\geqq d_1(u)\sigma+o(\sigma)\end{aligned}$$

as $\sigma\to 0_+$, with $d_1(u)>0$. Now observe that for some $\sigma_1\in(0,\alpha]$ and all $\sigma\in(0,\sigma_1]$, the foregoing estimate holds with

$$\left|\frac{o(\sigma)}{\sigma}\right| \leqq (1-\delta)d_1(u) \leqq (1-\delta)[s_N^{-1}\sigma^{-2}\|u+\sigma v_T-\phi(\sigma)\|^2+\langle P_{N^*}(-\nabla J(u)),\, v_T\rangle]. \qquad □$$

COROLLARY 1. *The fixed points of the* SGP *iteration* (1) *coincide with the stationary points of J in Ω.*

Proof. Suppose that v and σ are determined by (1). If u is stationary, then $\phi(\sigma)=u$ (see the proof of Lemma 3). Conversely, if u is not stationary, then $J(u)-J(\phi(\sigma))\geqq \delta d_1(u)\sigma>0$ and therefore $\phi(\sigma)\neq u$. □

Lemma 3 justifies the Bertsekas-Gafni steplength rule in the general setting of (1). No matter how the cones are constructed, the rule (1m) is well posed and strictly decreases J when u is nonstationary. Consequently, the sequences $\{J(u^i)\}$ produced by (1) are typically decreasing and must converge to $-\infty$ or some finite limit. In the latter case, it can be shown that the quantities $\sigma^i d(u^i)^2$ and $u^{i+1}-u^i$ must converge to zero. While these results alone do not ensure limit point stationarity or convergence of $\{u^i\}$ to any limit, they are important in the development of global and local convergence theories for (1)-(2) and (3)-(4).

LEMMA 4. *Assume that Ω is a closed convex set and that J is Fréchet differentiable in Ω. Let $\{u^i\}$ and $\{\sigma^i\}$ be generated by* (1). *Then $\{J(u^i)\}$ is nonincreasing and either*

$$\lim_{i\to\infty} J(u^i)=-\infty \tag{12a}$$

or

$$\lim_{i\to\infty} J(u^i)=\inf_i J(u^i)>-\infty \tag{12b}$$

and then

$$\lim_{i\to\infty} \sigma^i\|P_{N^{i*}}(-\nabla J(u^i))\|^2=0, \tag{12c}$$

$$\lim_{i\to\infty} (\sigma^i)^{-1}\|u^i-u^{i+1}\|^2=0, \tag{12d}$$

$$\lim_{i\to\infty} \|u^i-u^{i+1}\|^2=0, \tag{12e}$$

$$\lim_{i\to\infty} \sigma^i d(u^i)^2=0. \tag{12f}$$

Proof. Lemma 2 and (1m) imply that $\{J(u^i)\}$ is nonincreasing. If $\{J(u^i)\}$ is also bounded below, then (12b) holds and

$$\lim_{i\to\infty} (J(u^i)-J(u^{i+1}))=0.$$

The estimates (12c)-(12e) now follow easily from (1m), (10), and (11). To prove (12f), observe that (1j)-(1l), (7), and (9) yield

$$\begin{aligned} cd(u^i)&\leqq (s_N^i\sigma^i)^{-1}\|u^i-P_\Omega(u^i-s_N^i\sigma^i\nabla J(u^i))\|\\ &\leqq (\mu_2\sigma^i)^{-1}\|u^i-P_\Omega(u^i-s_N^i\sigma^i\nabla J(u^i))\|\\ &\leqq (\mu_2\sigma^i)^{-1}(\|u^i-u^{i+1}\|+(\mu_1+\mu_3)\sigma^i\|P_{N^{i*}}(-\nabla J(u^i))\|)\end{aligned}$$

with

$$c=\min\{1,(\alpha\mu_3)^{-1}\}>0.$$

Multiply both sides of this inequality by $\sqrt{\sigma^i}$, and (12f) will then follow from (12c)-(12d). □

3. Limit point stationarity. Lemma 4 falls short of establishing that every limit point of $\{u^i\}$ is stationary. This basic result will follow from (12f) if it can be shown that the steplengths σ^i associated with convergent subsequences of $\{u^i\}$ are bounded away from zero. Positive lower bounds for σ^i have indeed been demonstrated in [3] and [5] for unscaled GP algorithms (1) with $N=\mathcal{U}$; however, the proofs in question require locally Lipschitz continuous gradients ∇J, and do not extend readily to SGP iterations (1) with arbitrary closed convex cones $N \supset K(u)$. Reference [9] uses a different approach to establish limit point stationarity for unscaled GP processes (1) in polyhedral Ω when ∇J is merely continuous; with small modifications, the techniques in [9] will also prove the somewhat stronger result

$$\lim_{i\to\infty} d(u^i)=0$$

for unscaled GP when ∇J is uniformly continuous. Once again, however, the proof technique in [9] does not seem to work for (1) with arbitrary closed convex cones $N \supset K(u)$. It would appear that some further restrictions are needed on the cones N.

Proposition 2 of [12] establishes limit point stationarity for SGP sequences $\{u^i\}$ generated by (1)-(2) and the special class of closed convex cones

(13a) $$N^i = C^{i*} \supset C(u^i)^*,$$

where

(13b) $$C^i=\{w: \forall j\in \mathcal{J}^i, \langle w, a^j\rangle \leqq 0\}$$

and

(13c) $$\mathcal{J}^i \supset \mathcal{J}(u^i).$$

Nevertheless, with minor alterations the proof in [12] actually works for any closed convex cones N^i containing $C(u^i)^*$, and, in particular, for any closed *subspaces* $N^i \supset C(u^i)^*$. As explained below, this simple revision quickly yields a limit point stationarity theorem for the alternative SGP scheme (3)-(4) in polyhedral convex Ω; however, a more extensive proof modification is needed for (3)-(4) in nonpolyhedral sets (4a).

The following results sharpen the estimates in Lemmas 1 and 2 when the cones N satisfy (2).

LEMMA 5. *Let u belong to a closed convex set Ω with representation* (2a). *Fix $\varepsilon>0$ and put*

$$\mathcal{J}_\varepsilon(u)=\{j: \langle a^j, u\rangle - b^j \geqq -\|a^j\|\varepsilon\},$$
$$C_\varepsilon(u)=\{w: \langle a^j, w\rangle \leqq 0, \forall j\in\mathcal{J}_\varepsilon(u)\}.$$

Then for all $u'\in\Omega$, x and v in $\mathcal{U}$, and $\sigma\geqq 0$

$$\|u'-u\|<\varepsilon \Rightarrow K(u')\subset C_\varepsilon(u)^*$$

and

$$x\in C_\varepsilon(u) \quad \text{and} \quad \sigma\|v\|<\varepsilon \Rightarrow \langle x, \phi(\sigma)-u\rangle \geqq \langle x, v\rangle\sigma.$$

Proof. Suppose that $u'\in\Omega$ and $\|u'-u\|<\varepsilon$. Let w be a nonzero vector in the cone $C_\varepsilon(u)$, choose $h>0$ such that $0<h\|w\|\leqq \varepsilon-\|u'-u\|$, and put $v=u'+hw$. If $j\notin\mathcal{J}_\varepsilon(u)$, then

$$\begin{aligned}\langle a^j, v\rangle - b^j &= \langle a^j, u\rangle - b^j + \langle a^j, u'-u\rangle + \langle a^j, w\rangle h\\ &< \|a^j\|(-\varepsilon+\|u'-u\|+h\|w\|)\\ &\leqq 0.\end{aligned}$$

On the other hand, if $j \in \mathcal{J}_\varepsilon(u)$ then for any $h > 0$,

$$\langle a^j, v\rangle - b^j = \langle a^j, u'\rangle - b^j + \langle a^j, w\rangle h \leqq 0.$$

Consequently, $v \in \Omega$ and hence for all $w^* \in K(u')$,

$$\langle w^*, w\rangle = h^{-1}\langle w^*, v - u'\rangle \leqq 0.$$

Since this estimate holds for any vector w in $C_\varepsilon(u)$ it follows that $K(u') \subset C_\varepsilon(u)^*$. Therefore, in view of (5),

$$\begin{aligned} \varepsilon > \sigma\|v\| \geqq 0 &\Rightarrow K(\phi(\sigma)) \subset C_\varepsilon(u)^* \\ &\Rightarrow u + \sigma v - \phi(\sigma) \in C_\varepsilon(u)^* \\ &\Rightarrow \forall x \in C_\varepsilon(u), \quad \langle x, u + \sigma v - \phi(\sigma)\rangle \leqq 0. \qquad \square \end{aligned}$$

LEMMA 6. *Assume that Ω is a closed convex set with representation* (2a). *Fix u in Ω, suppose that J is Fréchet differentiable at u, and let v satisfy* (1b)–(1l) *and* (2). *Then for all $\sigma > 0$*

$$\varepsilon(u) > \sigma\|v\| \Rightarrow \langle \nabla J(u), u - \phi(\sigma)\rangle \geqq [s_N^{-1}\sigma^{-2}\|u + \sigma v_T - \phi(\sigma)\|^2 + \langle P_{N^*}(-\nabla J(u)), v_T\rangle]\sigma,$$

where the bracketed term on the right is bounded below by the quantity $d_1(u)$ in Lemma 2 when $\sigma \in (0, \alpha]$.

Proof. The lemma is proved by a repetition of the proof of Lemma 2, with Lemma 1 replaced by Lemma 5, and $\varepsilon = \varepsilon(u)$. □

Lemma 6 removes the restriction (13) imposed in Proposition 1 of [12] and produces a straightforward extension of the limit point stationarity result in Proposition 2 of [12].

THEOREM 1. *Assume that Ω is a closed convex set with representation* (2a). *Let J be continuously Fréchet differentiable in Ω and suppose that $\{u^i\}$ is generated by an* SGP *iteration* (1)–(2). *Then every subsequential limit of $\{u^i\}$ is stationary.*

Proof. Let Z_1 be an infinite set of positive integers and suppose that $u^i \to \bar{u}$ for $i \in Z_1$. Since Ω is closed and $d(\cdot)$ is continuous, $\bar{u}$ must lie in Ω and $d(u^i) \to d(\bar{u})$, $i \in Z_1$. As in the proof of Proposition 2 in [12], it will now be shown that $d(\bar{u}) > 0$ is impossible and that $\bar{u}$ must therefore be stationary.

Suppose that $d(\bar{u}) > 0$. Then by (12f)

$$\sigma^i \to 0, \qquad i \in Z_1.$$

In this case, the rule (1m) yields

$$J(u^i) - J(\phi^i) < \delta[(\beta^{-1}\sigma^i s_N^i)^{-1}\|u + \beta^{-1}\sigma^i v_T^i - \phi^i\|^2 + \langle P_{N^{i*}}(-\nabla J(u^i)), v_T^i\rangle \beta^{-1}\sigma^i] \tag{14a}$$

for $i \in Z_1$ and i sufficiently large, where

$$\phi^i = P_\Omega(u^i + \beta^{-1}\sigma^i v^i). \tag{14b}$$

Furthermore, since $\nabla J(\cdot)$ and $\varepsilon(\cdot)$ are continuous, (1j)–(1l) imply that for some $M > 0$,

$$\|v^i\| \leqq (\mu_1 + \mu_3)\|\nabla J(u^i)\| \leqq M, \qquad i \in Z_1, \tag{15}$$

$$\sigma^i\|v^i\| \to 0, \qquad i \in Z_1, \tag{16}$$

and

$$\beta^{-1}\sigma^i\|v\| < \varepsilon(u^i), \quad i \in Z_1, \quad i \text{ sufficiently large.}$$

In view of Lemma 6, inequality (5), and the mean value theorem, there are vectors $\xi^i \in \Omega$ on the line segment joining u^i and ϕ^i such that

$$\begin{aligned} J(u^i)-J(\phi^i) &\geqq \langle \nabla J(u^i), u^i-\phi^i\rangle - M\|\nabla J(\xi^i)-\nabla J(u^i)\|\beta^{-1}\sigma^i \\ &\geqq (\beta^{-1}\sigma^i s_N^i)^{-1}\|u^i+\beta^{-1}\sigma^i v_T^i-\phi^i\|^2 \\ &\qquad + \langle P_{N^{i*}}(-\nabla J(u^i)), v_T^i\rangle \beta^{-1}\sigma^i - M\|\nabla J(\xi^i)-\nabla J(u^i)\|\beta^{-1}\sigma^i, \end{aligned}$$

for $i \in Z_1$ and i sufficiently large, with

$$\|\xi^i-\phi^i\| \leqq \|u^i-\phi^i\| \leqq \sigma^i\|v^i\|.$$

Conditions (1j), (1k), inequality (10), and the foregoing estimates now yield

$$\begin{aligned} M\|\nabla J(\xi^i)-\nabla J(u^i)\| \geqq (1-\delta)[\mu_2(\beta^{-1}\sigma^i s_N^i)^{-2}\|u^i+\beta^{-1}\sigma^i v_T^i-\phi^i\|^2 \\ + \mu_0\|P_{N^{i*}}(-\nabla J(u^i))\|^2] \geqq 0 \end{aligned}$$

for $i \in Z_1$ and i sufficiently large, with

$$\|\nabla J(\xi^i)-\nabla J(u^i)\| \to 0, \qquad i \in Z_1.$$

Consequently,

$$(\beta^{-1}\sigma^i s_N^i)^{-1}\|u^i+\beta^{-1}\sigma^i v_T^i-\phi^i\| \to 0, \qquad i \in Z_1, \tag{17a}$$

$$P_{N^{i*}}(-\nabla J(u^i)) \to 0, \qquad i \in Z_1, \tag{17b}$$

and therefore

$$v_T^i \to 0, \qquad i \in Z_1,$$

because of (11). On the other hand, conditions (1j), (9), and the nonexpansive property for P_Ω give

$$\begin{aligned} d(u^i) &\leqq (\beta^{-1}\sigma^i s_N^i)^{-1}\|u^i - P_\Omega(u^i-\beta^{-1}\sigma^i s_N^i \nabla J(u^i))\| \\ &\leqq (\beta^{-1}\sigma^i s_N^i)^{-1}[\|u^i+\beta^{-1}\sigma^i v_T^i-\phi^i\| + \|\beta^{-1}\sigma^i v_T^i\| + \|\phi^i \\ &\qquad + \|\phi^i - P_\Omega(u^i-\beta^{-1}\sigma^i s_N^i\nabla J(u^i))\|] \\ &\leqq (\beta^{-1}\sigma^i s_N^i)^{-1}\|u^i+\beta^{-1}\sigma^i v_T^i-\phi^i\| + 2\mu_2^{-1}\|v_T^i\| + \|P_{N^{i*}}(-\nabla J(u^i))\| \end{aligned}$$

for $i \in Z_1$ and i sufficiently large, and so

$$\liminf_{\substack{i\to\infty \\ i\in Z_1}} (\beta^{-1}\sigma^i s_N^i)^{-1}\|u^i+\beta^{-1}\sigma^i v_T^i-\phi^i\| \geqq d(\bar{u}) > 0.$$

This contradiction proves that $d(\bar{u})>0$ is impossible. □

The following observation yields a corollary of Theorem 1 for SGP processes (3)-(4) in polyhedral convex sets.

Note 1. If Ω has a representation (4a) satisfying the constraint qualification (Q)(i), then the Farkas lemma implies that the cone $C(u)^*$ in (2) consists of zero and all nonnegative linear combinations of vectors in the set $\{a^j\}_{j\in\mathcal{J}(u)}$ with $\mathcal{J}(u)$ defined by (2d). Since the index set (2D) is no larger than its counterpart in (4e), it follows that $C(u)^*$ is contained in the closed subspace $N(u)$ in (4i). Hence for polyhedral convex Ω satisfying (Q)(i), the scheme (3)-(4) is subsumed by (1)-(2).

COROLLARY 1 OF THEOREM 1. *Assume that Ω is a polyhedral convex set with a representation* (4a) *that satisfies the constraint qualification* (Q)(i). *Let J be continuously Fréchet differentiable in Ω and suppose that $\{u^i\}$ is generated by an* SGP *iteration* (3)-(4). *Then every limit point of $\{u^i\}$ is stationary.*

Proof. The proof is immediate from Note 1 and Theorem 1. □

It has just been observed that (3)-(4) is a special case of (1)-(2) in polyhedral convex sets; moreover, it can be shown that if we choose $N = C(u)^*$ in (1)-(2), then the resulting SGP iteration is actually *equivalent* to some iteration (3)-(4) with $N = N(u)$ near nondegenerate stationary points $\bar{u}$ satisfying

$$-\nabla J(\bar{u}) \in \text{ri}\, K(\bar{u}); \tag{18}$$

i.e., the two constructions generate identical iterates near $\bar{u}$ even though $N(u) \neq C(u)^*$ in general. On the other hand, in nonpolyhedral convex sets (4a), $N(u)$ and $C(u)^*$ can be very different objects indeed, and (3)-(4) is neither a special case of (1)-(2), nor is it locally equivalent to (1)-(2). In particular, this means that limit point stationarity for (3)-(4) cannot be inferred from Theorem 1 in nonpolyhedral sets. Nevertheless, all limit points of (3)-(4) are stationary in closed convex sets (4a) satisfying (Q)(ii), and a proof can be based on the following alternatives to Lemmas 5 and 6.

The term $\phi(\sigma)$ appearing in Lemma 6 actually depends on u and v as well as σ, and for present purposes it is better to stress this dependence explicitly by writing $P_\Omega(u+\sigma v)$ in place of $\phi(\sigma)$.

LEMMA 7. *Let $\bar{u}$ be a nonstationary point for J in a closed convex set Ω with a representation* (4a) *that satisfies the constraint qualification* (Q)(ii). *Let $N(u)$ be defined by* (4) *at $u \in \Omega$, put $T(u) = N(u)^\perp$, and assume that the functions g_j in* (4a) *are continuously Fréchet differentiable at $\bar{u}$. Then for each $M > 0$ there are corresponding numbers $\Delta > 0$, $\rho > 0$, and $\kappa > 0$ such that for all $\sigma \in (0, \Delta]$, $u \in B(\bar{u}, \rho) \cap \Omega$, $x \in T(u)$, and $v \in B(0, M)$,*

$$\langle x, P_\Omega(u+\sigma v) - u\rangle \geqq \langle x, v\rangle \sigma + \|x\| \eta(\sigma, u, v)$$

with

$$\begin{aligned}\eta(\sigma, u, v) &= -\kappa \|\nabla g(P_\Omega(u+\sigma v)) - \nabla g(u)\|_2 \\ &\triangleq -\kappa \left(\sum_{j=1}^{m} \|\nabla g_j(P_\Omega(u+\sigma v)) - \nabla g_j(u)\|^2 \right)^{1/2}.\end{aligned}$$

Proof. (Q)(ii) and the Farkas lemma imply that each $u \in \Omega$, $K(u)$ consists of zero and all nonnegative linear combinations of vectors from the linearly independent set $\mathcal{G}_0(u)$. Furthermore, (5) implies that for all $\sigma \geqq 0$, $u \in \Omega$, and $v \in B(0, M)$

$$\|P_\Omega(u+\sigma v) - \bar{u}\| \leqq M\sigma + \|u - \bar{u}\|.$$

Since g_j and ∇g_j are continuous at $\bar{u}$, it follows that for some $\rho_1 > 0$, $\Delta_1 > 0$, $\gamma > 0$, and all $u \in B(\bar{u}, \rho_1) \cap \Omega$, $\sigma \in [0, \Delta_1]$, $v \in B(0, M)$, and $c \in \mathbb{R}^m$

$$\mathcal{J}_0(P_\Omega(u+\sigma v)) \subset \mathcal{J}_0(\bar{u}),$$

$$\{\nabla g_j(P_\Omega(u+\sigma v))\}_{j \in \mathcal{J}_0(\bar{u})} \quad \text{is linearly independent,}$$

and

$$\gamma \left\| \sum_{j \in \mathcal{J}_0(\bar{u})} c_j \nabla g_j(P_\Omega(u+\sigma v)) \right\| \geqq \left(\sum_{j \in \mathcal{J}_0(\bar{u})} c_j^2 \right)^{1/2}$$

Now write

$$\langle x, P_\Omega(u+\sigma v)-u\rangle = \langle x, v\rangle\sigma - \langle x, \psi(\sigma, u, v)\rangle$$

with

$$\psi(\sigma, u, v) = (u+\sigma v - P_\Omega(u+\sigma v)) \in K(P_\Omega(u+\sigma v))$$

and

$$\|\psi(\sigma, u, v)\| \leqq 2M\sigma.$$

For each $u \in B(\bar{u}, \rho_1)\cap\Omega$, $\sigma\in[0, \Delta_1]$, and $v \in B(0, M)$, there is a corresponding nonnegative $c \in \mathbb{R}^m$ such that

$$(j \notin \mathcal{J}_0(P_\Omega(u+\sigma v)) \Rightarrow c_j = 0), \qquad j = 1, \cdots, m,$$

$$\psi(\sigma, u, v) = \sum_{j=1}^{m} c_j \nabla g_j(P_\Omega(u+\sigma v)),$$

and

$$2\gamma M\sigma \geqq \left(\sum_{j=1}^{m} c_j^2\right)^{1/2},$$

and therefore

$$-\langle x, \psi(\sigma, u, v)\rangle \geqq \|x\|\eta(\sigma, u, v) - \left|\left\langle x, \sum_{j=1}^{m} c_j \nabla g_j(u)\right\rangle\right|$$

with

$$\eta(\sigma, u, v) = -2\gamma M\|\nabla g(P_\Omega(u+\sigma v)) - \nabla g(u)\|_2\sigma.$$

Now consider that for all $u \in \Omega$, $\sigma \geqq 0$, $v \in B(0, M)$, and $j \in \mathcal{J}_0(P_\Omega(u+\sigma v))$ there is a $\xi^j \in \Omega$ on the line segment between u and $P_\Omega(u+\sigma v)$ such that

$$g_j(u) \geqq -\|\nabla g_j(\xi^j)\|M\sigma$$

with

$$\|\xi^j - \bar{u}\| \leqq M\sigma + \|u-\bar{u}\|.$$

This follows at once from (5) and the mean value theorem. Since the functions $\varepsilon(\cdot)$ and $\|\nabla g_j(\cdot)\|$ are continuous and positive at $\bar{u}$ for $j \in \mathcal{J}_0(\bar{u})$, and since $\theta > 1$ in (4), it is now apparent that for some $\rho \in (0, \rho_1]$, $\Delta \in (0, \Delta_1]$, and all $u \in B(\bar{u}, \rho)\cap\Omega$, $\sigma \in [0, \Delta]$, $v \in B(0, M)$, and $j \in \mathcal{J}_0(P_\Omega(u+\sigma v))$,

$$\begin{aligned} g_j(u) &\geqq -\|\nabla g_j(u)\|\varepsilon(u) + (\tfrac{1}{2}\|\nabla g_j(u)\|\varepsilon(u) - \|\nabla g_j(\xi^j) - \nabla g_j(u)\|M\sigma) \\ &\geqq -\theta\|\nabla g_j(u)\|\varepsilon(u) \end{aligned}$$

and therefore

$$\mathcal{J}_0(P_\Omega(u+\sigma v)) \subset \mathcal{J}(u),$$

$$\sum_{j=1}^{m} c_j \nabla g_j(u) \in N(u),$$

and

$$\left\langle x, \sum_{j \in \mathcal{J}_0(P_\Omega(u+\sigma v))} c_j \nabla g_j(u)\right\rangle = 0$$

for all $x \in T(u)$. □

LEMMA 8. *Let $\bar{u}$ be a nonstationary point in a closed convex set Ω with a representation* (4a) *that satisfies the constraint qualification* (Q)(ii). *Assume that J and the functions g_j in* (4a) *are continuously Fréchet differentiable at $\bar{u}$ and suppose that v is determined by* (3b)-(3j) *and* (4) *at each $u \in \Omega$. Then for some $\rho > 0$, $\Delta > 0$, $\kappa > 0$, and all $u \in B(\bar{u}, \rho) \cap \Omega$ and $\sigma \in (0, \Delta]$,*

$$\langle \nabla J(u), u - P_\Omega(u + \sigma v) \rangle \geqq [s_N^{-1} \sigma^{-2} \| u + \sigma v_T - P_\Omega(u + \sigma v) \|^2 - \langle P_T \nabla J(u), v_T \rangle] \sigma \\ + s_N^{-1} \| x \| \eta(\sigma, u, v),$$

where

$$x = (v_T - s_N P_T \nabla J(u)) \in T,$$
$$\eta(\sigma, u, v) = -\kappa \| \nabla g(P_\Omega(u + \sigma v)) - \nabla g(u) \|_2 \sigma,$$

and the bracketed term on the right is bounded below by $d_1(u)$ in Lemma 2 *when $\sigma \in (0, \alpha]$ and $N^* = N^\perp = T$.*

Proof. Since $\nabla J(\cdot)$ is continuous at $\bar{u}$, and

$$\| v \| = \| -s_N P_N \nabla J(u) - S_T P_T \nabla J(u) \| \leqq (\mu_1 + \mu_3) \| \nabla J(u) \|,$$

it follows that $\| v \|$ is bounded above by some $M > 0$ when u is near $\bar{u}$ in Ω. Now repeat the proof of Lemma 2, with Lemma 1 replaced by Lemma 7. □

THEOREM 2. *Assume that Ω is a closed convex set with a representation* (4a) *that satisfies the constraint qualification* (Q)(ii). *Let J and the functions g_j in* (4a) *be continuously differentiable in Ω and suppose that $\{u^i\}$ is generated by an* SGP *iteration* (3)-(4). *Then every subsequential limit of $\{u^i\}$ is stationary.*

Proof. The proof scheme for Theorem 1 can be used here as well, with Lemma 8 taking the place of Lemma 6. Once again, if $u^i \to \bar{u}$, $i \in Z_1$, then $\bar{u}$ must lie in Ω. Moreover, if $d(\bar{u}) > 0$ then assertions (13)-(16) must hold as before, except that now $[T] = T = N^\perp = N^*$. Let M be the bound in (15). Then for all sufficiently large $i \in Z_1$, conditions (5), (13), Lemma 8, and the mean value theorem imply that for some $\xi^i \in \Omega$,

$$J(u^i) - J(\phi^i) \geqq \langle \nabla J(u^i), u^i - \phi^i \rangle - M \| \nabla J(\xi^i) - \nabla J(u^i) \| \beta^{-1} \sigma^i \\ - \kappa s_N^i \| x^i \| \, \| \nabla g(\phi^i) - \nabla g(u^i) \|_2 \beta^{-1} \sigma^i - M \| \nabla J(\xi^i) - \nabla J(u^i) \| \beta^{-1} \sigma^i$$

with

$$\phi^i = P_\Omega(u^i + \beta^{-1} \sigma^i v^i),$$
$$\| \xi^i - \phi^i \| \leqq \| u^i - \phi^i \| \leqq \sigma^i \| v^i \|,$$

and

$$\| x^i \| = \| v_T^i - s_N^i P_{T^i} \nabla J(u^i) \| \\ \leqq (\mu_1 + \mu_3) \| \nabla J(u^i) \| \\ \leqq M.$$

These estimates and (14) now produce

$$M \| \nabla J(\xi^i) - \nabla J(u^i) \| + \kappa \mu_3 \| \nabla g(\phi^i) - \nabla g(u^i) \|_2 \\ \geqq (1 - \delta) [\mu_2 (\beta^{-1} \sigma^i s_N^i)^{-2} \| u^i + \beta^{-1} \sigma^i v_T^i - \phi^i \|^2 + \mu_0 \| P_{T^i} \nabla J(u^i) \|^2] \geqq 0$$

for large $i \in Z_1$, with

$$\| \nabla J(\xi^i) - \nabla J(u^i) \| \to 0, \qquad i \in Z_1,$$

and

$$\|\nabla g(\phi^i)-\nabla g(u^i)\|_2 \to 0, \qquad i \in Z_1.$$

This leads to (17) as before, and from here on the proof is the same as the proof for Theorem 1. □

Note 2. Theorems 1 and 2 remain valid if the function $\varepsilon(\cdot)$ in (2) and (4) is replaced by any continuous measure of nonstationarity (i.e., any nonnegative continuous real function whose zeros coincide with the stationary points of J in Ω; in particular, this means that the parameter θ in (4) can be absorbed in $\varepsilon(\cdot)$ and is therefore superfluous in the present context. In fact, Theorem 1 continues to hold if $C(u)^*$ is replaced by any closed convex cone with the inclusion property expressed in Lemma 5 when $\varepsilon=\varepsilon(u)$, and $\varepsilon(\cdot)$ is a continuous measure of nonstationarity. However, local convergence proofs for (1)-(2) and (3)-(4) are more closely tied to the forms imposed on $\mathscr{I}(u)$, $C(u)^*$, and $N(u)$ in (2) and (4); this point is developed further in a sequel.

4. Global convergence in compact level sets. If all limit points of the sequence $\{u^i\}$ lie in a closed set $\mathscr{S}$, and if every subsequence of $\{u^i\}$ has a limit point, then it is easily seen that $\{u^i\}$ must converge to $\mathscr{S}$, i.e.,

$$\lim_{i\to\infty}\inf_{u\in\mathscr{S}}\|u^i-u\|=0.$$

Moreover, if $\mathscr{S}$ is a finite set and if

$$\lim_{i\to\infty}\|u^i-u^{i+1}\|=0,$$

then $\{u^i\}$ must converge to some vector in $\mathscr{S}$, i.e.,

$$\exists u\in\mathscr{S},\quad \lim_{i\to\infty}u^i=u.$$

These general observations and the results in §§ 2 and 3 produce a *global* convergence theorem for SGP iterations in compact sets, i.e., a convergence result that applies to all SGP sequences $\{u^i\}$ irrespective of their starting points u^1 in Ω.

THEOREM 3. *Let J be continuously Fréchet differentiable and assume that one of the following conditions holds*:

(i) Ω *is a closed convex set with representation* (2a), *and* $\{u^i\}$ *is generated by an* SGP *iteration* (1)-(2).

(ii) Ω *is a closed convex set with representation* (4a) *satisfying the constraint qualification* (Q), g_j *is continuously Fréchet differentiable for* $j=1,\cdots,m$, *and* $\{u^i\}$ *is generated by an* SGP *iteration* (3)-(4).

In addition, suppose that the level sets of J in Ω are compact. Then $\{u^i\}$ converges to the subset of stationary points

$$\mathscr{S}=\{u\colon d(u)=0 \textit{ and } J(u)\leqq J(u^1)\}$$

and

$$\lim_{i\to\infty}d(u^i)=0.$$

Furthermore, if $\mathscr{S}$ is a finite set, then $\{u^i\}$ converges to some vector in $\mathscr{S}$.

Proof. By a straightforward application of Lemma 4, Theorem 1 and its corollary, and Theorem 2, the theorem is proved. □

REFERENCES

[1] A. A. GOLDSTEIN, *Convex programming in Hilbert space*, Bull. Amer. Math. Soc., 70 (1964), pp. 709-710.

[2] E. S. LEVITIN AND B. T. POLYAK, *Constrained minimization problems*, USSR Comput. Math. Phys., 6 (1966), pp. 1-50.

[3] D. P. BERTSEKAS, *On the Goldstein-Levitin-Polyak gradient projection method*, in Proc. 1974 IEEE Conference on Decision and Control, Phoenix, AZ, pp. 47-52; IEEE Trans. Automat. Control, 10 (1976), pp. 174-184.

[4] G. P. MCCORMICK AND R. A. TAPIA, *The gradient projection method under mild differentiability conditions*, SIAM J. Control Optim., 10 (1972), pp. 93-98.

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

[6] ———, *On the convergence of projected gradient processes to singular attractors*, J. Optim. Theory Appl., 55 (1987), pp. 203-215.

[7] M. GAWANDE, *Projection algorithms for specially structured constrained minimization problems*, Ph.D. thesis, North Carolina State University, Raleigh, NC, 1986.

[8] M. GAWANDE AND J. C. DUNN, *Variable metric gradient projection processes in convex feasible sets defined by nonlinear inequalities*, Appl. Math. Optim., 17 (1988), pp. 103-119.

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

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

[11] D. P. BERTSEKAS, *Projected Newton methods for optimization problems with simple constraints*, SIAM J. Control Optim., 20 (1982), pp. 221-246.

[12] E. M. GAFNI AND D. P. BERTSEKAS, *Two-metric projection methods for constrained minimization*, SIAM J. Control Optim., 22 (1984), pp. 936-964.

[13] M. GAWANDE AND J. C. DUNN, *A projected Newton method in a Cartesian product of balls*, J. Optim. Theory Appl., 59 (1988), pp. 45-69.

[14] J. C. DUNN, *A projected Newton method for minimization problems with nonlinear inequality constraints*, Numer. Math., 53 (1988), pp. 377-409.

[15] ———, *Gradient projection methods for systems optimization problems*, in Control and Dynamic Systems, 29, C. T. Leondes, ed., Academic Press, Orlando, FL, 1988.

[16] J. E. DENNIS, JR. AND R. B. SCHNABLE, *Numerical methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[17] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, 1982.

[18] L. ARMIJO, *Minimization of functions having continuous partial derivatives*, Pacific J. Math., 16 (1966), pp. 1-3.

[19] A. A. GOLDSTEIN, *On steepest descent*, SIAM J. Control Optim., 3 (1965), pp. 147-151.

[20] ———, *On Newton's method*, Numer. Math., 7 (1965), pp. 391-393.

[21] B. D. CRAVEN AND J. J. KOLIHA, *Generalizations of Farkas' theorem*, SIAM J. Math. Anal., 8 (1977), pp. 983-997.

[22] S. J. WRIGHT, *Solution of discrete-time optimal control problems on parallel computers*, Parallel Comput., 16 (1990), pp. 221-238.

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

[24] E. H. ZARANTONELLO, *Projections on convex sets in Hilbert space, and spectral theory*, in Contributions to Nonlinear Functional Analysis, E. H. Zarantonello, ed., Academic Press, New York, 1971.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 5, pp. 1176–1215, September 1991

009

THE SENSITIVITY OF OPTIMAL CONTROL PROBLEMS TO TIME DELAY*

F. H. CLARKE† AND P. R. WOLENSKI‡

Abstract. A general class of time-delayed optimal control problems is studied. The goal is to characterize, in terms of the multipliers occurring in the necessary conditions, the rate of change of the value of the problem with respect to change in the delay. In particular, criteria for this dependence to be differentiable are given as well as a formula for the derivative. These appear to be the first results of this type.

Résumé. On étudie une classe générale de problèmes en contrôle optimal avec retard, le but étant de caractériser le taux de change de la valeur du problème par rapport au changement du retard, et ceci en terme des multiplicateurs figurant dans des conditions nécessaires d'optimalité. Comme cas particulier, on donne des critères impliquant que cette fonction soit différentiable, ainsi qu'une formule pour sa dérivée. Ces résultats semblent être les premiers de ce genre.

Key words. delayed systems, hereditary, retarded, nonsmooth analysis

AMS(MOS) subject classifications. 49E99; secondary 34K35

1. Introduction.

Differential-difference equations and time delay (other terminology: lag, retarded, hereditary) optimal control problems have long been recognized as important models for real-life phenomena. However, explicit solutions to such problems are difficult to represent in even the simplest cases. If the time delay is very small, it may be preferable to ignore the delay and assume that the system is instantaneous. What error accrues from this simplification? The present paper addresses the issue of the sensitivity of the value function in time delay optimal control problems to small changes in the delay parameter. In the case of differential-difference equations, an explicit formula is derived for the derivative of the trajectory endpoint with respect to the delay.

Consider the following time delay optimal control problem:

$$\min J(x,u) \quad \text{over}$$

$x(\cdot)$ absolutely continuous on $[a,b]$, $u(\cdot)$ measurable on $[a,b]$ satisfying

$$\begin{aligned} &\dot{x}(t) = \phi(t, x(t), x(t-\Delta), u(t)) \quad \text{a.e. } t \in [a,b], \\ &u(t) \in U(t) \quad \text{a.e. } t \in [a,b], \qquad\qquad (1.1)\\ &x(t) = c(t) \quad \text{for} \quad t \in [a-\Delta, a]. \end{aligned}$$

Here J is a Bolza-type functional, and the data is given under appropriate assumptions (see §3). The delay Δ is constant but is a parameter of the problem. Let $V(\Delta)$ be the optimal value. The first main result (Theorem 3.1 below) contains bounds for certain

*Received by the editors August 13, 1990; accepted for publication October 18, 1990. This work was supported in part by the Natural Sciences and Engineering Research Council (Canada) and Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (le Fonds FCAR, Quebec).

†Centre de Recherches Mathématiques, Université de Montréal, POB 6128-A, Montréal, Québec, H3C 3J7 Canada.

‡Department of Mathematics, Louisiana State University, Baton Rouge, Louisiana 70803.

Dini derivatives of V. These estimates contain a term of the form

$$\int_a^b \langle -\dot{q}(t), \dot{x}(t-\Delta)\rangle \,\mathrm{dt}, \tag{1.2}$$

where $x(\cdot)$ is an optimal trajectory of (1.1), and $q(\cdot)$ is a "piece" of the adjoint variable associated with $x(\cdot)$ arising from the maximum principle. The integral (1.2) is of a type introduced here for the first time, for it couples the derivatives of primal and dual variables. It is not known whether (1.2) has geometrical significance.

It is more delicate to prove the opposite bounds on the Dini derivatives. Essentially, we need information on the limiting behavior of (1.2) under certain perturbations of the data, but this is difficult to obtain without introducing further assumptions. Two approaches are taken. The first hypothesizes "pointwise convergence of optimal controls" (see (H8) in §3) for which the conclusion (Theorem 3.2) is that one-sided derivatives of V in Δ actually exist and equal the respective bounds found in Theorem 3.1. The extra hypothesis (H8) is unsatisfactory due to its impractical nature, but on the other hand, we know of no example for which it fails. The second approach introduces a smoothness assumption on the Hamiltonian (see (H9) and (H10) in §6), and the derivatives of V are described (Theorem 6.1) in terms of Hamiltonian multipliers. The main results are reformulated in Theorem 7.1 for the case in which an endpoint constraint $x(b) \in C$ is added to the basic problem (1.1).

The basic theory of differential-difference equations is covered by Bellman and Cooke [BC2], Halanay [H1], and Driver [D2]. The major research trend is to treat time delay problems under the general framework of functional differential equations (see Hale [H2]). The state space in the latter is infinite-dimensional, which is appropriate in generalizing many topics from ordinary differential equation theory. Our approach is to revert back to a finite-dimensional viewpoint, which is natural under the circumstances. The problems under consideration here have constant delay, which could be zero. For simplicity, only one delay component is taken account of, but multiple delay components could be easily handled by the same methods.

A consequence of our results identifies the derivative of the trajectory endpoint of an ordinary differential equation (o.d.e.) whenever a time delay is introduced into the equation (see §8). Not much attention has been given to the behavior of systems under a change in the delay. Sugiyama [S] has shown that the trajectory depends continuously in the sup norm with respect to the delay parameter. Bellman and Cooke [BC1] obtained this result earlier with linear systems. The Sugiyama result follows directly from ours. Driver [D1] has derived conditions in which delay equations exhibit the same asymptotic properties as those without delay.

An introduction to time delay optimal control problems can be found in Oğuztöreli [O], Warga [W1], and Manitius [M1]. The time-delay version of the maximum principle was proven by Kharatishvili [K]. See [P, §27]. Clarke and Watkins [CW] derived a nonsmooth version of the maximum principle in the form of a Hamiltonian inclusion, a result crucial to the treatment of endpoint constraints in §7. Necessary conditions with nonsmooth data are also given by Warga [W2], [W3].

As already mentioned, our basic theory is applicable for only constant delay parameters. It would be of interest to extend our results to allow for delays dependent on time, state, and/or control. Maximum principles are available for such problems. See Banks [B] and Manitius [M2]. The impediment to generalizing our methods in this direction occurs in Lemma 4.1, where we must differentiate a function of the form $\Delta \longrightarrow \phi(x(t-\Delta))$, where ϕ is C^1 and $x(\cdot)$ is merely absolutely continuous. If Δ is a function of other parameters, such a (Banach space) differentiation is difficult to perform and utilize.

The methodology in the proof is that of nonsmooth analysis. We first bound the upper Dini derivative by a certain generalized directional derivative, which in turn is represented as the support function of the generalized gradient (Clarke [C1]). Since the generalized gradients are characterized as the closed convex hull of weak limits of proximal subgradients, the bound is seen to be a limit of more manageable quantities. The line of reasoning to this point has become standard [C1], [C2]. The novel approach in this paper is in the problem formulation and the consequent proximal subgradients that come under consideration. In brief, we will be led to consider proximal subgradients of an augmented value function $V(\Delta, \alpha)$, where $(\Delta, \alpha) \in \mathbb{R}^+ \times L^2[a,b]$. The variable Δ is the delay parameter, and α is a perturbation of the dynamic equation. The variable α does not appear in the original problem but is introduced to monitor the Δ-dependence. In general, proximal subgradients in a product space split into proximal subgradients of the components (not true of generalized gradients). This is of paramount interest to us, because the subgradient with respect to α can be calculated and represented as a dual variable (the nondelay case is in Clarke [C2]), whence the subgradient with respect to Δ procures a specific representation.

The outline of the paper is as follows. Section 2 contains preliminaries (problem formulation, maximum principle, proximal analysis); §3 contains the basic hypotheses and the statements of two main results; the proofs are given simultaneously in §4; §5 weakens the smoothness hypothesis on the endpoint cost function ℓ; §6 offers a Hamiltonian approach to a lower bound; the main results are formulated in §7 to include an endpoint constraint; and, in §8, the results are illustrated by three examples.

2. Preliminaries.

This section contains basic definitions and reviews some preliminary results. The analysis of estimating the Dini derivatives of the value function will require introducing certain modified and perturbed problems. We begin by describing the basic problem and these alterations of it.

Throughout the paper, $c(\cdot)$ is a fixed Lipschitz continuous function of order $||\dot{c}||_\infty$ on $[a-\overline{\Delta}, a]$, where $\overline{\Delta} > 0$. In all choices of Δ below, it is understood, if not explicitly stated, that $0 \leq \Delta < \overline{\Delta}$. Suppose we are given

$$
\begin{aligned}
\ell &: [0,\overline{\Delta}) \times \mathbb{R}^n \longrightarrow \mathbb{R}^1 \\
L &: [a,b] \times [0,\overline{\Delta}) \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^1 \\
\phi &: [a,b] \times [0,\overline{\Delta}) \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n \\
U &: [a,b] \rightrightarrows \mathbb{R}^m.
\end{aligned}
$$

(The double arrow denotes that U is a set-valued map.) Precise assumptions on the data will be given in §3. The functional J is defined by

$$
\begin{aligned}
&J : [0,\overline{\Delta}) \times AC[a,b] \times \mathscr{M}[a,b] \longrightarrow \mathbb{R}^1 \cup \{\pm\infty\}, \\
&J(\Delta, x(\cdot), u(\cdot)) = \ell(\Delta, x(b)) + \int_a^b L(t, \Delta, x(t), x(t-\Delta), u(t))\, dt,
\end{aligned}
\tag{2.1}
$$

where $x(t-\Delta)$ is set equal to $c(t-\Delta)$ whenever $t - \Delta \leq a$. Here and elsewhere, $AC_n[a,b]$ denotes the absolutely continuous functions on $[a,b]$ into $\mathbb{R}^n$ (the subscript n will be dropped if the dimension is clear from the context), and $\mathscr{M}[a,b]$ is the set of measurable functions from $[a,b]$ into $\mathbb{R}^m$. If $u(\cdot) \in \mathscr{M}[a,b]$ satisfies $u(t) \in U(t)$ almost everywhere $t \in [a,b]$, then $u(\cdot)$ is called admissible.

The problem $P(\Delta)$ is defined as

$$\min J(\Delta, x, u) \text{ over } (x,u) \text{ satisfying}$$

$$\begin{aligned} &\dot{x}(t) = \phi(t, \Delta, x(t), x(t-\Delta), u(t)) \quad \text{a.e. } t \in [a,b], \\ &u(t) \in U(t) \quad \text{a.e. } t \in [a,b], \\ &x(t) = c(t) \quad \text{for} \quad t \in [a-\Delta, a]. \end{aligned} \tag{2.2}$$

Suppose that $v(\cdot) \in L^2_m[a,b]$ (:= the square integrable functions on $[a,b]$ with values in $\mathbb{R}^m$; again the subscript m will be dropped whenever the dimension is clear from the context). The problem $P_v(\Delta)$ is defined in the same manner as $P(\Delta)$ except that the term $||u-v||_2^2$ is added to the objective, where $|| \cdot ||_2$ denotes the $L^2[a,b]$ norm.

Suppose that $\alpha(\cdot) \in L^2_n[a,b]$. The problem $P(\Delta, \alpha)$ is the same as $P(\Delta)$ except that now the dynamic equation $\dot{x} = \phi$ is replaced by the perturbed equation

$$\dot{x}(t) = \phi(t, \Delta, x(t), x(t-\Delta), u(t)) + \alpha(t).$$

The problem $P_v(\Delta, \alpha)$ is defined by incorporating both alterations of the last two paragraphs.

The optimal value of $P(\Delta)$ is denoted by $V(\Delta)$. Similarly $V_v(\Delta)$, $V(\Delta, \alpha)$, $V_v(\Delta, \alpha)$ denote optimal values of $P_v(\Delta)$, $P(\Delta, \alpha)$, $P_v(\Delta, \alpha)$, respectively.

Given $u(\cdot) \in \mathscr{M}[a,b]$ admissible, $\alpha(\cdot) \in L^2_n[a,b]$, and $\Delta \in [0, \overline{\Delta})$, our assumptions on ϕ will imply that there exists a unique $x(\cdot) \in AC[a,b]$ satisfying $\dot{x}(t) = \phi(t, \Delta, x(t), x(t-\Delta), u(t)) + \alpha(t)$ almost everywhere $t \in [a,b]$, and $x(t-\Delta) = c(t-\Delta)$ whenever $t - \Delta \leq a$. This trajectory will be denoted by $x^{u,\Delta}_{\alpha}$. If $\alpha = 0$, we will simply write $x^{u,\Delta}$. The optimal solutions of $P(\Delta)$, that, is the *controls* that are minimizers, will be denoted by $\Sigma(\Delta)$. Similarly, we use the notation $\Sigma_v(\Delta)$, $\Sigma(\Delta, \alpha)$, and $\Sigma_v(\Delta, \alpha)$.

The (pseudo-) Hamiltonian $\mathscr{H}$ is defined by

$$\mathscr{H} : [a,b] \times [0, \overline{\Delta}) \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^n \longrightarrow \mathbb{R}^1$$

$$\mathscr{H}(t, \Delta, x, y, u, p) = \langle p, \phi(t, \Delta, x, y, u) \rangle - L(t, \Delta, x, y, u).$$

The preparation for and statements of the main results can be given more succinctly by utilizing $\mathscr{H}$. The classical maximum principle as it applies to the problem $P(\Delta)$ is due to Kharatishvili (see [K]) and says the following.

MAXIMUM PRINCIPLE. *Suppose that* $u(\cdot) \in \Sigma(\Delta)$ *and set* $x(\cdot) = x^{u,\Delta}(\cdot)$. *Then there exist absolutely continuous functions* $p(\cdot)$ *and* $q(\cdot)$ *from* $[a,b]$ *into* $\mathbb{R}^n$ *so that*

(i) $-p(b) = \nabla_x \ell(\Delta, x(b)), \quad q(b) = 0,$

(ii)

$$-\dot{p}(t) = \frac{\partial}{\partial x}\mathscr{H}^*(t, \Delta, x(t), x(t-\Delta), u(t), p(t) + q(t+\Delta)) \quad a.e. \;\; t \in [a,b],$$

$$-\dot{q}(t) = \frac{\partial}{\partial y}\mathscr{H}^*(t, \Delta, x(t), x(t-\Delta), u(t), p(t) + q(t+\Delta)) \quad a.e. \;\; t \in [a,b],$$

and

(iii) *for almost all* $t \in [a,b]$,
$\max \mathscr{H}(t, \Delta, x(t), x(t-\Delta), u, p(t) + q(t+\Delta))$ *over* $u \in U(t)$
occurs at $u = u(t)$.

In the above, the star (*) denotes transpose and, whenever $t+\Delta \geq b$, then $q(t+\Delta)$ is set equal to zero. Part (i) is the tranversality condition, (ii) contains the adjoint equations, and (iii) is the maximum condition.

Let P be a problem as described above, and Σ its set of optimal solutions. Fix $u(\cdot) \in \Sigma$. The maximum principle is applicable under the assumptions invoked in §3. Also, the pair (p, q) (referred to as a Pontryagin multiplier) is unique, and for notational convenience we set $M(u, P) = (p, q)$. Note that this notation and terminology is retained in the case when $P = P(0)$. In this special case, where $\Delta = 0$ and the problem is without delay, multipliers that arise from the usual maximum principle are broken into two pieces to become multipliers in our sense (this is illustrated in Example C of §8).

Techniques of proximal analysis play a prominent role in the analysis below. We give here a concise compilation of the material that will be used. The reader may consult [C1] or [C3] for a complete introduction to these ideas.

Suppose $X \subseteq \mathscr{X}$, where $\mathscr{X}$ is a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$. Suppose $f : \mathscr{X} \to \mathbb{R}^1$ is locally Lipschitz near X and $x_0 \in X$. An element $\beta \in \mathscr{X}$ is a proximal subgradient of f at x_0 if there exists $\sigma > 0$ and a neighborhood Y of x_0 so that for each $x \in Y$, we have

$$f(x) \geq f(x_0) + \langle \beta, x - x_0 \rangle - \sigma ||x - x_0||^2. \tag{2.3}$$

The presubdifferential $\hat{\partial} f(x_0)$ and the generalized gradient $\partial f(x_0)$ of f at x_0 are defined as the weak limits of proximal subgradients and the closed convex hull of these, respectively:

$$\begin{aligned} \hat{\partial} f(x_0) = \{ \beta \in \mathscr{X} :\ & \text{there exist sequences } \{x_i\}_{i=1}^{\infty} \text{ and } \{\beta_i\}_{i=1}^{\infty} \text{ in } \mathscr{X} \\ & \text{so that } x_i \to x_0 \text{ strongly, } \beta_i \to \beta \text{ weakly , and } \beta_i \text{ is a proximal} \\ & \text{subgradient of } f \text{ at } x_i \}, \end{aligned} \tag{2.4}$$

and

$$\partial f(x_0) = \operatorname{cl} \operatorname{co} \hat{\partial} f(x_0). \tag{2.5}$$

In view of the extension to Hilbert spaces of Clarke's proximal normal formula (see Borwein and Strojwas [BS] or Loewen [L]), the definition of $\partial f(x_0)$ coincides with the usual definition that is given in [C1]. In particular, if x_0 is in X and $\omega \in \mathscr{X}$, the following holds:

$$\limsup_{\substack{x \to x_0 \\ h \downarrow 0}} \frac{f(x + h\omega) - f(x)}{h} = \sup \{ \langle \omega, \beta \rangle : \beta \in \partial f(x_0) \}. \tag{2.6}$$

The left side of (2.6) is denoted by $f^0(x_0; \omega)$. Note that $\partial f(x_0)$ on the right side of (2.6) can be replaced by $\hat{\partial} f(x_0)$. The set-valued map $x \to \partial f(x)$ is upper semicontinuous in the strong to weak topology, which means that if $x_i \to x_0$ strongly and $\beta_i \to \beta_0$ weakly with $\beta_i \in \partial f(x_i)$, then $\beta_0 \in \partial f(x_0)$. The same is true for $\hat{\partial} f(x)$.

The only Hilbert spaces of interest here will be $\mathbb{R}^n$ and $\mathbb{R}^1 \times L_n^2[a, b]$. The notation $\langle \cdot, \cdot \rangle$ will be used to denote the inner product on $\mathbb{R}^1 \times L_n^2[a, b]$ as well as the usual inner product on $\mathbb{R}^n$. This should cause no confusion. Hence for (Δ, α), $(\Delta', \alpha') \in \mathbb{R}^1 \times L_n^2[a, b]$, we write

$$\langle (\Delta, \alpha), (\Delta', \alpha') \rangle = \Delta \cdot \Delta' + \int_a^b \langle \alpha(t), \alpha'(t) \rangle \, \mathrm{dt}.$$

3. The main result. Recall the problem formulation $P(\Delta)$. The basic hypotheses on the data ℓ, L, ϕ, and U are the following. The notation L_Δ (respectively, L_1, L_2) is used to denote the derivative of $\Delta \to L(t, \Delta, x, y, u)$ (respectively, $x \to L(t, \Delta, x, y, u)$, $y \to L(t, \Delta, x, y, u)$), and similarly for ϕ_Δ, ϕ_1, ϕ_2.

(H1) $\ell : [0, \overline{\Delta}) \times \mathbb{R}^n \longrightarrow \mathbb{R}^1$ is C^1.

(H2) $L : [a, b] \times [0, \overline{\Delta}) \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^1$ is measurable in the first variable whenever the other variables are held fixed. For almost all $t \in [a, b]$, L is continuous in (Δ, x, y, u), and L_Δ, L_1, L_2 exist and are continuous in (Δ, x, y, u).

(H3) $\phi : [a, b] \times [0, \overline{\Delta}) \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ is measurable in the first variable whenever the other variables are held fixed. For almost all $t \in [a, b]$, ϕ is continuous in (Δ, x, y, u), and ϕ_Δ, ϕ_1, ϕ_2 exist and are continuous in (Δ, x, y, u).

(H4) $U : [a, b] \rightrightarrows \mathbb{R}^m$ is measurable with nonempty compact values, and $\sup\{|u| : u \in U(t)\} \leq \gamma_1(t)$ almost everywhere $t \in [a, b]$ for some $\gamma_1 \in L^2[a, b]$.

(H5) ϕ and $[a, b]$ are such that for each $R > 0$, there exists $r_R > 0$ so that whenever $\alpha \in L^2[a, b]$ with $||\alpha||_2 \leq R$, $u(\cdot) \in \mathscr{M}[a, b]$ with $u(t) \in U(t)$ almost everywhere $t \in [a, b]$, and $\Delta \in [0, \overline{\Delta})$, the solution $x(\cdot)$ that satisfies $\dot{x}(t) = \phi(t, \Delta, x(t), x(t - \Delta), u(t)) + \alpha(t)$ almost everywhere $t \in [a, b]$ and $x(t) = c(t)$ on $[a - \Delta, a]$ also satisfies $||x|| \leq r_R$, where $|| \cdot ||$ denotes the sup norm on $[a - \Delta, b]$.

(H6) For each $r > 0$, there exist $\sigma_r(\cdot) \in L^2[a, b]$ and $\gamma_r(\cdot) \in L^1[a, b]$ so that $|\phi| + |\phi_\Delta| + |\phi_2| + |L_2| \leq \sigma_r(t)$, $|\phi_1| + |L| + |L_\Delta| + |L_1| \leq \gamma_r(t)$ almost everywhere $t \in [a, b]$ whenever the arguments (Δ, x, y, u) satisfy $\Delta \in [0, \overline{\Delta})$, $|x| \leq r$, $|y| \leq r$, $u \in U(t)$.

(H7) For each $\Delta \in [0, \overline{\Delta})$, $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$, and almost all $t \in [a, b]$, the set $\{(\phi(t, \Delta, x, y, u), L(t, \Delta, x, y, u) + \delta) : u \in U(t), \delta \geq 0\}$ is a closed convex subset of $\mathbb{R}^n \times \mathbb{R}^1$.

Remarks. (i) Assumption (H5) is satisfied, for example, if $\sup\{|\phi(t, \Delta, x, y, u)| : u \in U(t)\} \leq \gamma_2(t) + \gamma_3(t)(|x| + |y|)$ almost everywhere $t \in [a, b]$, where γ_2, $\gamma_3 \in L^1[a, b]$. The proof is an exercise in applying Gronwall's inequality and is left to the reader.

(ii) The growth conditions in (H6) and regularity conditions in (H3) on ϕ combine to guarantee that the dynamics in each problem $P(\Delta, \alpha)$ are satisfied by at least one $x(\cdot)$. The convexity assumption (H7) assures that the solution set Σ is nonempty for each problem P.

(iii) As mentioned in §2, techniques of nonsmooth analysis play a prominent role, and so it may be somewhat surprising that we hypothesize C^1 data rather than Lipschitz. The only difficulty in weakening the assumptions in this manner occurs in Lemma 4.1, where a chain rule is used with one of the functions absolutely continuous and not known to be Lipschitz. Hence if L or ϕ are Lipschitz but not C^1, nonsmooth chain rules (see, for example, [C1]) apparently do not apply. It is not clear how to overcome this obstacle. No difficulty arises, however, if the endpoint cost ℓ is assumed merely Lipschitz in the state variable, and we will do this in §5.

To facilitate the statements of the main results, we define the following functional E:

$$E\colon\ [0,\overline{\Delta}) \times AC_n[a,b] \times \mathscr{M}[a,b] \times AC_{2n}[a,b] \longrightarrow \mathbb{R}^1$$

$$\text{(3.1)}\quad E(\Delta,x,v,(p,q)) := \frac{\partial}{\partial\Delta}\ell(\Delta,x(b)) - \int_a^b \left[\frac{\partial}{\partial\Delta}\mathscr{H}(t,\Delta,x(t),x(t-\Delta),v(t),p(t)+q(t+\Delta)) + \langle\dot{q}(t),\dot{x}(t-\Delta)\rangle\right] \mathrm{dt}.$$

If $t-\Delta \le a$, then $x(t-\Delta) := c(t-\Delta)$, and if $t+\Delta \ge b$, then $q(t+\Delta) := 0$. The first main result follows.

THEOREM 3.1. *Assume that* (H1)-(H7) *hold, and let* $\Delta_0 \in [0,\overline{\Delta})$. *Then*

(a)

$$\limsup_{h\searrow 0} \frac{V(\Delta_0+h)-V(\Delta_0)}{h} \le \inf_{v\in\Sigma(\Delta_0)} E\Big(\Delta_0, x^{v,\Delta_0}, v, M(v,P(\Delta_0))\Big);$$

(b) *If* $\Delta_0 > 0$, *then*

$$\liminf_{h\nearrow 0} \frac{V(\Delta_0+h)-V(\Delta_0)}{h} \ge \sup_{v\in\Sigma(\Delta_0)} E\Big(\Delta_0, x^{v,\Delta_0}, v, M(v,P(\Delta_0))\Big).$$

The proof will be given in the next section. Note that we avoid negative arguments in discussing V, since the problem $P(\Delta)$ is not even defined for $\Delta < 0$.

Theorem 3.1 provides estimates on two Dini derivatives of $V(\Delta)$. Subtleties arise in attempting to apply the same proof technique to obtain similar estimates on the complementary Dini derivatives. As we will see, the complications surface in taking limits of terms that have the same form as the integral term in (3.1). We next address this issue in further detail, which will serve to motivate the added assumption in Theorem 3.2.

Let $\alpha \in L^2[a,b]$, $v_h \in \Sigma(\Delta+h,\alpha)$, $x_h(\cdot) = x_\alpha^{v_h,\Delta+h}(\cdot)$, $(p_h,q_h) = M(v_h,P(\Delta+h))$, and define $w_h(\cdot)$ by

$$w_h(t) = -\int_a^t \frac{\partial}{\partial\Delta}\mathscr{H}(s,\Delta+h,x_h(s),x_h(s-\Delta-h),v_h(s),p_h(s)+q_h(s+\Delta+h))\,\mathrm{ds}.$$

Then the integral term in (3.1) with this data is

$$\text{(3.2)}\qquad w_h(b) + \int_a^b \langle -\dot{q}_h(t), \dot{x}_h(t-\Delta-h)\rangle\,\mathrm{dt}.$$

We are interested in the behavior of (3.2) as $h \searrow 0$. The natural inclination is to apply the Dunford–Pettis criterion to $\{\dot{q}_h\}$ and $\{\dot{x}_h\}$ and obtain absolutely continuous functions $q(\cdot)$ and $x(\cdot)$, for which subsequences of $\{\dot{q}_h\}$ and $\{\dot{x}_h\}$ converge to $\dot{q}$ and $\dot{x}$ weakly. However, as the first and most serious objection, this is inadequate in determining the behavior of the integral in (3.2). Second, the $w_h(b)$ term depends on the optimal controls $v_h(\cdot)$, and it does not follow that $w_h(b)$ will converge as $h \to 0$ to an appropriate limit (an ad hoc and unsatisfactory convexity hypothesis could be introduced on the function $(\partial/\partial\Delta)\mathscr{H}(t,\Delta,x,y,\cdot,p)$ to mollify this objection). A third obstacle arises in determining the limiting function $q(\cdot)$. Is it a piece of the multiplier? The answer is "not necessarily," for the Pontryagin multipliers $M(v,P(\Delta))$ do not have sufficient closure properties under perturbations of the data to assure this.

One approach that overcomes the three above-mentioned pitfalls is to hypothesize "convergence of optimal controls." Another, more readily verifiable, approach will be given in §6. Consider

(H8) Fix $\Delta_0 \in (0, \overline{\Delta})$ (respectively, $\Delta_0 = 0$). For each pair of sequences $\{\alpha_i\}_{i=1}^{\infty} \subseteq L_n^2[a,b]$ and $\{h_i\}_{i=1}^{\infty}$ with $\alpha_i \to 0$ and $h_i \to 0$ (respectively, $h_i \searrow 0$) as $i \to \infty$, there exists $\{v_i(\cdot)\}_{i=1}^{\infty}$ with $v_i \in \Sigma(\Delta_0 + h_i, \alpha_i)$ so that a subsequence of $\{v_i(\cdot)\}$ converges almost everywhere on $[a,b]$ to some $v(\cdot) \in \mathscr{M}[a,b]$.

In a practical situation, (H8) may be very difficult to verify. On the other hand, it seems that producing an example for which it fails is not trivial. If $v(\cdot) \in \Sigma(\Delta)$, then we will show that (H8) holds under (Hl) - (H7) for the problem $P_v(\Delta)$. It is this fact that allows for the estimates in Theorem 3.1 to be obtained without further assumptions. Unfortunately, the penalization technique is inapplicable for estimating the opposite bounds.

The following asserts the existence of one-sided derivatives of V.

THEOREM 3.2. *Let* $\Delta_0 \in [0, \overline{\Delta})$, *and suppose* (Hl)–(H8) *hold. Then*

(a) $$\lim_{h \searrow 0} \frac{V(\Delta_0 + h) - V(\Delta_0)}{h} = \inf_{v \in \Sigma(\Delta_0)} E\Big(\Delta_0, x^{v,\Delta_0}, v, M(v, P(\Delta_0))\Big);$$

(b) *If* $\Delta_0 > 0$, *then*

$$\lim_{h \nearrow 0} \frac{V(\Delta_0 + h) - V(\Delta_0)}{h} = \sup_{v \in \Sigma(\Delta_0)} E\Big(\Delta_0, x^{v,\Delta_0}, v, M(v, P(\Delta_0))\Big).$$

The proof will be given in §4, simultaneously with that of Theorem 3.1. A criterion for two-sided differentiability can be immediately deduced from Theorem 3.2. We state this in the following corollary.

COROLLARY 3.3. *Let* $\Delta_0 \in (0, \overline{\Delta})$, *and suppose* (Hl)–(H8) *hold. Suppose further that* $\Sigma(\Delta_0)$ *consists of a single element* v. *Then* V *is differentiable at* Δ_0, *with derivative*

$$\frac{d}{d\Delta} V(\Delta_0) = E\Big(\Delta_0, x^{v,\Delta_0}, v, M(v, P(\Delta_0))\Big).$$

Note that, as a consequence of the expressions derived in Theorem 3.2 for the one-sided derivatives, we conclude that V will not be differentiable, in general, when the solution to the problem is not unique.

4. Proof Of Theorems 3.1 and 3.2. The proofs of Theorem 3.1 (a) and 3.2 (a) are broken into seven steps. The basic outline of the proof is this. It is shown first (Step 1) that the value function is locally Lipschitz in (Δ, α) on $[0, \overline{\Delta}) \times L_n^2[a,b]$. Next, estimates of the lim sup in Theorem 3.1 (a) and the analogous limit in Theorem 3.2 (a) are obtained (Step 2) by introducing a modified problem and applying the formula (2.4). This estimate further leads to introducing an auxiliary optimization problem arising from the definition of proximal subgradients. Since a solution of this auxiliary problem is known, necessary optimality conditions are applied (Step 3) to derive the main identity. The necessary conditions in the maximum principle then represent (Step 4) a proximal subgradient as a multiplier. The previous material is brought together (Step 5) into upper and lower estimates involving rather complicated limits. Finally, extensive limiting arguments (Steps 6 and 7) finish the proof of Theorem 3.1 (a). The proof of Theorem 3.2 (a) will be completed by altering slightly the final limiting arguments. The proofs of both (b) parts require only minor modifications in Step 2.

Step 1. It will be first shown that $V(\Delta, \alpha)$ is locally Lipschitz in (Δ, α) on $[0, \overline{\Delta}) \times L^2[a,b]$. Fix $\Delta \in [0, \overline{\Delta})$ and $R > 0$. Set $r := r_R$, where r_R is as in (H5). Suppose that $\alpha_1, \alpha_2 \in L^2[a,b]$ are so that $||\alpha_1||_2$, $||\alpha_2||_2 \leq R$. Let $u(\cdot) \in \Sigma(P(\Delta, \alpha_1))$, and let $x_1(\cdot) = x_{\alpha_1}^{u,\Delta}(\cdot)$ and $x_2(\cdot) = x_{\alpha_2}^{u,\Delta}$. Recall that this means

$$\dot{x}_1(t) = \phi(t, \Delta, x_1(t), x_1(t-\Delta), u(t)) + \alpha_1(t) \quad \text{a.e. } t \in [a,b],$$

$$x_1(t) = c(t) \quad \text{for} \quad t \in [a-\Delta, a],$$

and $x_2(\cdot)$ is a similar solution with α_1 replaced by α_2. We have for each $t \in [a,b]$,

$$\begin{aligned} |x_1(t) - x_2(t)| &\leq \int_a^t |\dot{x}_1(s) - \dot{x}_2(s)| \, \mathrm{ds} \\ &\leq \int_a^t |\alpha_1(s) - \alpha_2(s)| \, \mathrm{ds} + \int_a^t (\gamma_r(s) + \sigma_r(s)) |x_1(s) - x_2(s)| \, \mathrm{ds} \qquad (4.1) \\ &\leq (b-a)^{1/2} ||\alpha_1 - \alpha_2||_2 + \int_a^t (\gamma_r(s) + \sigma_r(s)) |x_1(s) - x_2(s)| \, \mathrm{ds}, \end{aligned}$$

where $\gamma_r(\cdot) \in L^1[a,b]$ and $\sigma_r(\cdot) \in L^2[a,b]$ are chosen as in (H6). From (4.1) and Gronwall's lemma, we deduce that

$$|x_1(t) - x_2(t)| \leq k_1 ||\alpha_1 - \alpha_2||_2 \tag{4.2}$$

for each $t \in [a,b]$, where $k_1 = (b-a)^{1/2}(1 + (||\gamma_r||_1 + ||\sigma_r||_1) \exp(||\gamma_r||_1 + ||\sigma_r||_1))$. Now (x_2, u) satisfies the dynamics of $P(\Delta, \alpha_2)$, so we have

$$\begin{aligned} V(\Delta, \alpha_2) &\leq J(\Delta, x_2, u) \\ &= \ell(\Delta, x_2(b)) + \int_a^b L(t, \Delta, x_2(t), x_2(t-\Delta), u(t)) \, \mathrm{dt} \\ &\leq \ell(\Delta, x_1(b)) + \int_a^b L(t, \Delta, x_1(t), x_1(t-\Delta), u(t)) \, \mathrm{dt} \qquad (4.3) \\ &\quad + k_2 \, |x_1(b) - x_2(b)| + \int_a^b (\gamma_r(t) + \sigma_r(t)) |x_1(t) - x_2(t)| \, \mathrm{dt}, \end{aligned}$$

where $k_2 = \sup\{\nabla_\xi \ell(\Delta, \xi) : ||\xi|| \leq r\}$. The last inequality is an immediate consequence of the mean value theorem, (H1) and (H6). Recall that $u \in \Sigma(P(\Delta, \alpha_1))$, so from (4.2) and (4.3) we have

$$V(\Delta, \alpha_2) \leq V(\Delta, \alpha_1) + k_3 ||\alpha_1 - \alpha_2||_2, \tag{4.4}$$

where $k_3 = k_1(k_2 + ||\gamma_r||_1 + ||\sigma_r||_1)$.

By retracing the above steps with α_1 and α_2 interchanged, we conclude from (4.4) that $V(\Delta, \cdot)$ is locally Lipschitz. Moreover, note that the Lipschitz constant does not depend on the choice of $\Delta \in [0, \overline{\Delta})$.

We now turn to the Lipschitz dependence with respect to Δ. Let $R > 0$, $0 < \Delta_1 < \Delta_2 < \overline{\Delta}$, and $\alpha \in L^2[a,b]$ with $||\alpha||_2 \leq R$. Set $r = r_R$. Let $u \in \Sigma(P(\Delta, \alpha))$, and let $x_1(\cdot) = x_\alpha^{u,\Delta_1}(\cdot)$ and $x_2(\cdot) = x_\alpha^{u,\Delta_2}(\cdot)$.

We have for each $t \in [a, b]$

$$\tag{4.5}\begin{aligned}|x_1(t) - x_2(t)| &\le \int_a^t |\dot{x}_1(s) - \dot{x}_2(s)|\,\mathrm{ds}\\ &= \int_a^t |\phi(s, \Delta_1, x_1(s), x_1(s - \Delta_1), u(s))\\ &\qquad\qquad - \phi(s, \Delta_2, x_2(s), x_2(s - \Delta_2), u(s))|\,\mathrm{ds}\\ &\le \int_a^t \{\sigma_r(s)(\Delta_2 - \Delta_1) + \sigma_r(s)|x_1(s - \Delta_1) - x_2(s - \Delta_2)|\\ &\qquad\qquad + \gamma_r(s)|x_1(s) - x_2(s)\}\ \mathrm{ds},\end{aligned}$$

where $\sigma_r(\cdot)$ and $\gamma_r(\cdot)$ are as in (H6).

Consider the second term in the integrand on the right side of (4.5). For each $s \in [a, t]$, we have

$$\tag{4.6}\begin{aligned}|x_1(s - \Delta_1) - x_2(s - \Delta_2)| &\le \int_{s-\Delta_2}^{s-\Delta_1} |\dot{x}_1(s')|\,\mathrm{ds}' + |x_1(s - \Delta_2) - x_2(s - \Delta_2)|\\ &\le \int_{s-\Delta_2}^{s-\Delta_1} (\sigma_r(s') + \alpha(s'))\,\mathrm{ds}' + |x_1(s - \Delta_2) - x_2(s - \Delta_2)|,\end{aligned}$$

where we set $\sigma_r(s) = \dot{c}(s)$ and $\alpha(s) = 0$ if $s < a$. Next, we use Hölder's inequality twice:

$$\tag{4.7}\begin{aligned}&\int_a^t \sigma_r(s) \int_{s-\Delta_2}^{s-\Delta_1} (\sigma_r(s') + \alpha(s'))\mathrm{ds}'\mathrm{ds}\\ &\le \|\sigma_r\|_2 \left\{ \int_a^t \left[\int_{s-\Delta_2}^{s-\Delta_1} (\sigma_r(s') + \alpha(s'))\mathrm{ds}' \right]^2 \mathrm{ds} \right\}^{1/2}\\ &\le \|\sigma_r\|_2(\Delta_2 - \Delta_1)^{1/2} \left\{ \int_a^t \int_{s-\Delta_2}^{s-\Delta_1} |\sigma_r(s') + \alpha(s')|^2\mathrm{ds}'\mathrm{ds} \right\}^{1/2}.\end{aligned}$$

Fubini's theorem applied to the double integral on the right-hand side of (4.7) gives

$$\begin{aligned}\int_a^t \int_{s-\Delta_2}^{s-\Delta_1} |\sigma_r(s') + \alpha(s')|^2\mathrm{ds}'\mathrm{ds} &\le \int_{a-\Delta_2}^{t-\Delta_1} \int_{s'-\Delta_1}^{s'-\Delta_2} |\sigma_r(s') + \alpha(s')|^2\mathrm{ds}\mathrm{ds}'\\ &\le (\Delta_2 - \Delta_1)\|\sigma_r + \alpha\|_2^2.\end{aligned}$$

Inserting this into (4.7) yields the estimate

$$\tag{4.8}\int_a^t \sigma_r(s) \int_{s-\Delta_2}^{s-\Delta_1} (\sigma_r(s') + \alpha(s'))\mathrm{ds}'\mathrm{ds} \le \|\sigma_r\|_2(\|\sigma_r\|_2 + \|\alpha\|_2)(\Delta_2 - \Delta_1).$$

Now we multiply (4.6) by $\sigma_r(s)$ and integrate from a to t. In light of (4.8), we obtain

$$\tag{4.9}\begin{aligned}&\int_a^t \sigma_r(s)|x_1(s - \Delta_1) - x_2(s - \Delta_2)|\mathrm{ds}\\ &\le \|\sigma_r\|_2(\|\sigma_r\|_2 + \|\alpha\|_2)(\Delta_2 - \Delta_2) + \int_a^t \sigma_r(s)|x_1(s - \Delta_2) - x_2(s - \Delta_2)|\mathrm{ds}.\end{aligned}$$

Plugging (4.9) into (4.5), we have

$$|x_1(t) - x_2(t)| \leq k_4(\Delta_2 - \Delta_1) + \int_a^t (\sigma_r(s + \Delta_2) + \gamma_r(s))|x_1(s) - x_2(s)|\mathrm{ds},$$

where $k_4 = \|\sigma_r\|_1 + \|\sigma_r\|_2(\|\sigma_r\|_2 + R)$, and $\sigma_r(s + \Delta_2)$ is set equal to zero if $s + \Delta_2$ exceeds b. Another application of Gronwall's lemma yields the inequality

$$|x_1(t) - x_2(t)| \leq k_5|\Delta_2 - \Delta_1|, \tag{4.10}$$

with $k_5 = k_4(1 + (\|\sigma_r\|_2 + \|\gamma_r\|_1) \exp\,(\|\sigma_r\|_1 + \|\gamma_r\|_1))$.

Analogously, similar reasoning shows that estimate (4.10) holds whenever $0 \leq \Delta_2 < \Delta_1 < \overline{\Delta}$.

The argument that derived (4.3) from (4.2) can be utilized here with only minor modifications, and the analogue to (4.4) is

$$|V(\Delta_2, \alpha) - V(\Delta_1, \alpha)| \leq k_6|\Delta_2 - \Delta_1| \tag{4.11}$$

for some constant k_6 that is independent of $\alpha \in L^2[a, b]$, provided $||\alpha||_2 \leq R$.

Finally, let $R > 0$ and set $k = \max\{k_3, k_6\}$ where k_3 and k_6 are as in (4.4) and (4.11). For each $\Delta_1, \Delta_2 \in [0, \overline{\Delta})$ and $\alpha_1, \alpha_2 \in L^2[a, b]$ with $||\alpha_1||_2$, $||\alpha_2||_2 \leq R$, we have

$$\begin{aligned}|V(\Delta_1, \alpha_1) - V(\Delta_2, \alpha_2)| &\leq |V(\Delta_1, \alpha_1) - V(\Delta_2, \alpha_1)| + |V(\Delta_2, \alpha_1) - V(\Delta_2, \alpha_2)| \\ &\leq k(|\Delta_1 - \Delta_2| + ||\alpha_1 - \alpha_2||_2).\end{aligned}$$

This is the statement that V is locally Lipschitz on $[0, \overline{\Delta}) \times L^2[a, b]$. □

Unless noted otherwise, $\Delta_0 \in (0, \overline{\Delta})$ is fixed. The case $\Delta_0 = 0$ will be handled separately at the end.

Step 2. Fix $v \in \Sigma(\Delta_0)$. Since $V_v(\Delta) \geq V(\Delta)$ for all $\Delta \in [0, \overline{\Delta})$ and $V_v(\Delta_0) = V(\Delta_0)$, we have

$$\begin{aligned}\limsup_{h\searrow 0} \frac{V(\Delta_0 + h) - V(\Delta_0)}{h} &\leq \limsup_{h\searrow 0} \frac{V_v(\Delta_0 + h) - V_v(\Delta_0)}{h} \\ &\leq \limsup_{\substack{h\searrow 0 \\ \Delta\to\Delta_0 \\ \alpha\to 0}} \frac{V_v(\Delta + h, \alpha) - V_v(\Delta, \alpha)}{h} \\ &= V_v^0((\Delta_0, 0); (1, 0)).\end{aligned}$$

Formula (2.6) is applicable because V_v is locally Lipschitz by Step 1. Hence the last above term equals

$$\sup\{\mu : (\mu, \beta) \in \partial V_v(\Delta_0, 0)\},$$

where the subgradient is taken with respect to both variables $(\Delta, \alpha) \in \mathbb{R}^+ \times L^2[a, b]$. This, in turn, can be expressed in terms of weak limits via (2.4) and (2.5), and since $v \in \Sigma(\Delta_0)$ is arbitrary, we obtain the following estimate:

$$\limsup_{h\searrow 0} \frac{V(\Delta_0 + h) - V(\Delta_0)}{h} \leq \inf_{v\in\Sigma(\Delta_0)} \sup \mu, \tag{4.12}$$

where the sup is taken over sequences $\{(\Delta_i, \alpha_i)\}_{i=1}^{\infty}$, $\{(\mu_i, \beta_i)\}_{i=1}^{\infty}$ of $\mathbb{R}^1 \times L^2[a, b]$ such that $(\Delta_i, \alpha_i) \to (\Delta_0, 0)$ strongly, $(\mu_i, \beta_i) \to (\mu, \beta)$ weakly (for some $\beta \in L^2[a, b]$), and (μ_i, β_i) is a proximal subgradient of V_v at (Δ_i, α_i).

A lower bound can also be obtained in a similar fashion. We have

$$\begin{aligned}\liminf_{h\searrow 0}\frac{V(\Delta_0+h)-V(\Delta_0)}{h} &\geq -\limsup_{\substack{h\searrow 0\\ \Delta\to\Delta_0\\ \alpha\to 0}}\frac{V(\Delta-h,\alpha)-V(\Delta,\alpha)}{h}\\ &= -V^0((\Delta_0,0);(-1,0))\\ &= \inf\{\mu:(\mu,\beta)\in\partial V(\Delta_0,0)\}.\end{aligned}$$

In a manner analogous to the above, we obtain the lower estimate

$$\liminf_{h\searrow 0}\frac{V(\Delta_0+h)-V(\Delta_0)}{h}\geq \inf\mu, \tag{4.13}$$

where the inf is taken over sequences $\{(\Delta_i,\alpha_i)\}_{i=1}^{\infty}$, $\{(\mu_i,\beta_i)\}_{i=1}^{\infty}$ of $\mathbb{R}^1\times L^2[a,b]$ such that $(\Delta_i,\alpha_i)\to(\Delta_0,0)$ strongly, $(\mu_i,\beta_i)\to(\mu,\beta)$ weakly (for some $\beta\in L^2[a,b]$), and (μ_i,β_i) is a proximal subgradient of V at (Δ_i,α_i).

Step 3. The bounds (4.12) and (4.13) introduce proximal subgradients and allow for the exploitation of the proximal subgradient inequality (2.3). The next lemma contains the fundamental identity. In this step and the next, we will take the lemmas with the object in mind to proving the estimates in Theorem 3.1. Similar statements are valid if the subscript "v" is dropped throughout, and these latter versions are required for the proof of Theorem 3.2.

LEMMA 4.1. *Suppose that* $(\Delta,\alpha)\in(0,\overline{\Delta})\times L^2[a,b]$, $v\in L^2[a,b]$, *and* $u\in\Sigma_v(\Delta,\alpha)$. *If* $(\mu,\beta)\in\mathbb{R}^1\times L^2[a,b]$ *is a proximal subgradient of* V_v *at* (Δ,α), *then*

$$\mu=\frac{\partial}{\partial\Delta}\ell(\Delta,x(b))+\int_a^b\{L_\Delta(\cdot)+\langle\beta(t),\phi_\Delta(\cdot)\rangle+\langle\phi_2^*(\cdot)(-\beta(t))-L_2^*(\cdot),\dot{x}(t-\Delta)\rangle\}\,dt,$$

where $x(\cdot)=x_\alpha^{u,\Delta}$, *and in the integrand, the functions with* $(\cdot)$ *are evaluated at* $(t,\Delta,x(t),x(t-\Delta),u(t))$.

Proof of Lemma 4.1. By the definition of proximal subgradients, there exists $\sigma>0$ so that for all $(\Delta',\alpha')\in(0,\overline{\Delta})\times L^2[a,b]$ near (Δ,α) in norm, we have

$$V_v(\Delta',\alpha')\geq V_v(\Delta,\alpha)+\langle(\mu,\beta),(\Delta'-\Delta,\alpha'-\alpha)\rangle-\sigma\|(\Delta'-\Delta,\alpha'-\alpha)\|^2. \tag{4.14}$$

If $y(\cdot)\in AC[a-\overline{\Delta},b]$ with $\dot{y}(\cdot)\in L^2[a,b]$ and $y(t)=c(t)$ on $[a-\Delta',a]$, and $\nu(\cdot)\in\mathscr{M}[a,b]$ with $\nu(t)\in U(t)$ almost everywhere $t\in[a,b]$, then $V_v(\Delta',\alpha')\leq J(\Delta',y,\nu)+\|\nu-v\|_2^2$, where $\alpha'(t):=\dot{y}(t)-\phi(t,\Delta,y(t),y(t-\Delta'),\nu(t))$. Substituting this choice of variables into (4.14), we can write

$$\begin{aligned}J(\Delta',y,\nu)+\|\nu-v\|_2^2-\mu\Delta'+\int_a^b\langle\beta(t),\phi(t,\Delta,y(t),y(t-\Delta'),\nu(t))-\dot{y}(t)\rangle\,dt\\ +\sigma\|(\Delta'-\Delta,\alpha'-\alpha)\|^2\geq V_v(\Delta,\alpha)-\langle(\mu,\beta),(\Delta,\alpha)\rangle.\end{aligned} \tag{4.15}$$

Now since $u\in\Sigma_v(\Delta,\alpha)$, the left-hand side of (4.15) has a local minimum in the variables (Δ',y,ν) at (Δ,x,u). Substitute (x,u) for (y,ν) in the left side of (4.15), and take the derivative with respect to Δ' at Δ. The result is zero, and the bounds on the integrands assumed on (H6) allow for differentiating under the integral. We obtain

$$\text{(4.16)}\quad \frac{\partial}{\partial\Delta}J(\Delta,x,u)-\mu+\int_a^b\langle\beta(t),\phi_\Delta(t,\Delta,x(t),x(t-\Delta),u(t)) \\ +\phi_2(t,\Delta,x(t),x(t-\Delta),u(t))(-\dot{x}(t-\Delta)\rangle\,\mathrm{dt}=0.$$

Recall that

$$J(\Delta,x,u)=\ell(\Delta,x(b))+\int_a^b L(t,\Delta,x(t),x(t-\Delta),u(t))\,\mathrm{dt},$$

and hence

(4.17)

$$\frac{\partial}{\partial\Delta}J(\Delta,x,u)=\frac{\partial}{\partial\Delta}\ell(\Delta,x(b))+\int_a^b L_\Delta(t,\Delta,x(t),x(t-\Delta),u(t))\,\mathrm{dt} \\ +\int_a^b\langle L_2^*(t,\Delta,x(t),x(t-\Delta),u(t)),-\dot{x}(t-\Delta)\rangle\,\mathrm{dt}.$$

Substituting (4.17) into (4.16) and rearranging terms gives

$$\mu=\frac{\partial}{\partial\Delta}\ell(\Delta,x(b)) \\ +\int_a^b\{L_\Delta(\cdot)+\langle\beta(t),\phi_\Delta(\cdot)\rangle+\langle\phi_2^*(\cdot)(-\beta(t))-L_2^*(\cdot),\dot{x}(t-\Delta)\rangle\}\,\mathrm{dt},$$

which is the assertion of the lemma. □

Step 4. The proximal subgradient inequality (4.15) can also be used to obtain information on the nature of $\beta(\cdot)$. The following lemma is an extension of Clarke [C2, Thm. 2.1] to problems with delay.

LEMMA 4.2. *In the notation of Lemma* 4.1, *we have* $-\beta(t)=p(t)+q(t+\Delta)$, *where* $(p,q)=M(u,P_v(\Delta,\alpha))$.

Proof. Replace Δ' by Δ in (4.15). Then it immediately follows from (4.15) that (x,u) is a local optimal solution to the optimal control problem

$$\min\ \ell(\Delta,y(b))+\int_a^b\Big\{L(t,\Delta,y(t),y(t-\Delta),\nu(t))+|\nu(t)-v(t)|^2 \\ +\langle\beta(t),\phi(t,\Delta,y(t),y(t-\Delta),\nu(t))-\dot{y}(t)\rangle \\ +\sigma|\dot{y}(t)-\phi(t,\Delta,y(t),y(t-\Delta),\nu(t))-\dot{x}(t) \\ +\phi(t,\Delta,x(t),x(t-\Delta),u(t))|^2\Big\}\,\mathrm{dt}$$

over possible (y,ν), and where $\dot{y}(t)$ is viewed as an additional control variable z without constraint. The maximum principle as given in §2 is applicable and asserts that there exist absolutely continuous functions $p(\cdot)$ and $q(\cdot)$ from $[a,b]$ to $\mathbb{R}^n$ such that

(i) $$-p(b) = \nabla_x \ell(\Delta, x(b)), \quad q(b) = 0,$$

(ii) $$-\dot{p}(t) = -L_1^*(t, \Delta, x(t), x(t-\Delta), u(t)) - \phi_1^*(t, \Delta, x(t), x(t-\Delta), u(t))\beta(t),$$ a.e. $t \in [a, b]$,

$$-\dot{q}(t) = -L_2^*(t, \Delta, x(t), x(t-\Delta), u(t)) - \phi_2^*(t, \Delta, x(t), x(t-\Delta), u(t))\beta(t),$$ a.e. $t \in [a, b]$,

and

(iii)
$$\begin{aligned}\max\{&\langle p(t) + q(t+\Delta), z\rangle - L(t, \Delta, x(t), x(t-\Delta), \nu) - |\nu - v(t)|^2\\ &-\langle \beta(t), \phi(t, \Delta, x(t), x(t-\Delta), \nu) - z\rangle\\ &-\sigma|z - \phi(t, \Delta, x(t), x(t-\Delta), \nu) - \dot{x}(t) + \phi(t, \Delta, x(t), x(t-\Delta), u(t))|^2\}\end{aligned}$$

over $\nu \in U(t)$, $z \in \mathbb{R}^n$ occurs at $\nu = u(t)$, $z = \dot{x}(t)$.

In the term maximized in (iii), substitute $u(t)$ for ν and differentiate with respect to z. Since this derivative is zero at $z = \dot{x}(t)$, it follows that $-\beta(t) = p(t) + q(t + \Delta)$. By substituting this into (ii), it becomes clear that the adjoint equations of the maximum principle applied to $(u, P_v(\Delta, \alpha))$ hold for (p, q). To conclude that $(p, q) = M(u, P_v(\Delta, \alpha))$ we must show that the maximum condition holds.

Again from (iii), we have that

$$\begin{aligned}(4.18)\quad \max\Bigg\{&\langle p(t) + q(t+\Delta), \phi(t, \Delta, x(t), x(t-\Delta), \nu)\rangle\\ &- L(t, \Delta, x(t), x(t-\Delta), \nu) - |\nu - v(t)|^2\\ &- \sigma|\phi(t, \Delta, x(t), x(t-\Delta), \nu) - \phi(t, \Delta, x(t), x(t-\Delta), u(t))|^2\Bigg\}\end{aligned}$$

over $\nu \in U(t)$ is achieved at $u(t) \in U(t)$. Let $S \subseteq \mathbb{R}^n \times \mathbb{R}^1$ be defined by

$$S := \{(-\phi(t, \Delta, x(t), x(t-\Delta), \nu), L(t, \Delta, x(t), x(t-\Delta), \nu) + |\nu - v(t)|^2 + \delta) : \delta \geq 0,\ \nu \in U(t)\},$$

and let $s_0 \in S$ be chosen as

$$s_0 = (-\phi(t, \Delta, x(t), x(t-\Delta), u(t)), L(t, \Delta, x(t), x(t-\Delta), u(t)) + |u(t) - v(t)|^2).$$

It follows from (4.18) that

$$\min\{\langle (p(t) + q(t+\Delta), 1), s\rangle + \sigma|s - s_0|^2\}$$

over $s \in S$ is attained at $s = s_0$. Now from the convexity hypothesis (H7), we have that S is convex. So from elementary convex analysis $\min\langle (p(t) + q(t+\Delta), 1), s\rangle$ over $s \in S$ is attained at $s = s_0$. Reverting back to the original notation, this means that the max of

$$\{\langle p(t) + q(t+\Delta), \phi(t, \Delta, x(t), x(t-\Delta), \nu)\rangle - L(t, \Delta, x(t), x(t-\Delta), \nu) - |\nu - v(t)|^2\}$$

over $\nu \in U(t)$ is achieved at $\nu = u(t)$. This is the maximum condition that finishes the demonstration that $(p, q) = M(u, P_v(\Delta, \alpha))$. □

Step 5. We now apply the results of Steps 3 and 4 to the proximal subgradients that arise in Step 2. The outcome of doing so is summarized in the following.

LEMMA 4.3. *We have*

(a)

$$\limsup_{h\searrow 0}\frac{V(\Delta_0+h)-V(\Delta_0)}{h}\le\inf_{v\in\Sigma(\Delta_0)}\ \limsup_{(\Delta,\alpha)}\ \inf_{u\in\Sigma_v(\Delta,\alpha)} E\Big(\Delta,x_\alpha^{u,\Delta},u,M(u,P_v(\Delta,\alpha))\Big);$$

(b)

$$\liminf_{h\searrow 0}\frac{V(\Delta_0+h)-V(\Delta_0)}{h}\ge\liminf_{(\Delta,\alpha)}\ \sup_{u\in\Sigma(\Delta,\alpha)} E\Big(\Delta,x_\alpha^{u,\Delta},u,M(u,P(\Delta,\alpha))\Big).$$

In (a), *the* lim sup *is taken over* $(\Delta,\alpha)\to(\Delta_0,0)$ *in* $\mathbb{R}^1\times L_n^2[a,b]$ *such that* V_v *has a proximal subgradient at* (Δ,α). *In* (b), *the* lim inf *is taken over* $(\Delta,\alpha)\to(\Delta_0,0)$ *in* $\mathbb{R}^1\times L_n^2[a,b]$ *such that* V *has a proximal subgradient at* (Δ,α).

Proof. (a) Recall estimate (4.12). Let $v\in\Sigma(\Delta_0)$ and $\{(\Delta_i,\alpha_i)\}_{i=1}^\infty$, $\{(\mu_i,\beta_i)\}_{i=1}^\infty$ be sequences of which the sup is taken over in (4.12). Pick any $u_i\in\Sigma_v(\Delta_i,\alpha_i)$ and set $x_i(\cdot)=x_{\alpha_i}^{u_i,\Delta_i}(\cdot)$. In light of Lemma 4.2, we may write $-\beta_i(t)=p_i(t)+q_i(t+\Delta_i)$ where $(p_i,q_i)=M(u_i,P_v(\Delta_i,\alpha_i))$. Observe that $q_i(\cdot)$ satisfies

$$-\dot q_i(t)=\frac{\partial}{\partial y}\mathscr{H}^*(t,\Delta_i,x_i(t),x_i(t-\Delta_i),u_i(t),p_i(t)+q_i(t+\Delta_i))\quad\text{a.e. } t\in[a,b],$$

and hence from Lemma 4.1 we conclude that

(4.19)

$$\begin{aligned}\mu_i=\frac{\partial}{\partial\Delta}\ell(\Delta_i,x_i(b))-\int_a^b\bigg[\frac{\partial}{\partial\Delta}\mathscr{H}\Big(t,\Delta_i,x_i(t),x_i(t-\Delta_i),u_i(t),p_i(t)+q_i(t+\Delta_i)\Big)\\ +\langle\dot q_i(t),\dot x_i(t-\Delta_i)\rangle\bigg]\,\mathrm{dt}\\ =E\Big(\Delta_i,x_i,u_i,M(u_i,P_v(\Delta_i,\alpha_i))\Big).\end{aligned}$$

Now (4.12) says that the right upper Dini derivative is bounded above by $\limsup_{i\to\infty}\mu_i$, and we have seen that each μ_i is represented as in (4.19). By then taking the inf over $v\in\Sigma(\Delta_0)$, it follows that (a) holds.

(b) The proof of (b) resembles that of (a). Estimate (4.13) gives a lower bound of the right lower Dini derivative of V at Δ_0 in the form $\liminf_{i\to\infty}\mu_i$, where μ_i has the representation (4.19) for some $u_i\in\Sigma(\Delta_i,\alpha_i)$. (Here we need to utilize the versions of Lemmas 4.1 and 4.2 that are without the subscript "v.") This is one of the choices the lim inf on the right side of (b) is taken over, hence estimate (b) is valid. □

The rest of the proofs of Theorems 3.1 (a) and 3.2 (a) now consists of calculating the lim sup and lim inf that occur in the right-hand sides of Lemma 4.3. Our proof will actually show that for fixed $v\in\Sigma(\Delta_0)$, the lim sup in the right-hand side of Lemma 4.3 (a) exists as a limit. Toward this end, we next show in effect that for $v\in\Sigma(\Delta_0)$, the problem $P_v(\Delta_0)$ satisfies (H8). In evaluating the lim inf in (b), we will not need this lemma, but rather will invoke (H8) directly.

LEMMA 4.4. *Let* $v\in\Sigma(\Delta_0)$, *and suppose* $\{h_i\}_{i=1}^\infty\subseteq\mathbb{R}^1$ *and* $\{\alpha_i\}_{i=1}^\infty\subseteq L^2[a,b]$ *are sequences such that* $h_i\to 0$ *and* $||\alpha_i||_2\to 0$ *as* $i\to\infty$. *Let* $u_i\in\Sigma_v(\Delta_i,\alpha_i)$. *Then a subsequence of* $\{u_i\}$ (that is not relabeled) *satisfies* $u_i(t)\to v(t)$ *almost everywhere* $t\in[a,b]$.

Proof. Note that $V_v(\Delta, \alpha)$ is continuous in (Δ, α) (Step 1). Hence

$$V(\Delta_0) = V_v(\Delta_0, 0) = \lim_{i\to\infty} V_v(\Delta_i, \alpha_i)$$

$$= \lim_{i\to\infty} \left\{ J(\Delta_i, x_i, u_i) + \|u_i - v\|_2^2 \right\}. \tag{4.20}$$

Suppose $\|u_i - v\|_2^2 \not\to 0$ as $i \to \infty$. Passing to a subsequence if necessary, assume $\|u_i - v\|_2^2 \geq \delta > 0$ for all i. Now $V(\Delta, \alpha)$ is also continuous in (Δ, α), and so

$$\begin{aligned} V(\Delta_0) = V(\Delta_0, 0) &= \lim_{i\to\infty} V(\Delta_i, \alpha_i) \\ &\leq \liminf_{i\to\infty} J(\Delta_i, x_i, u_i) \\ &\leq V(\Delta_0) - \delta. \end{aligned}$$

The last inequality follows from (4.20) and the assumption $\|u_i - v\|_2^2 \geq \delta$. This is a contradiction, so we must have $\|u_i - v\|_2^2 \to 0$, whence it follows that a subsequence satisfies $u_i(t) \to v(t)$ almost everywhere $t \in [a, b]$. □

Step 6. In this step, we show that the almost everywhere convergence of optimal controls induces nice convergence properties of the associated trajectories and multipliers.

We first focus attention on the lim sup in Lemma 4.3 (a). Fix $v(\cdot) \in \Sigma(\Delta_0)$. Let $\{\Delta_i\}_{i=1}^{\infty} \subseteq [0, \overline{\Delta})$, $\{\alpha_i\}_{i=1}^{\infty} \subseteq L^2[a, b]$ be such that $\Delta_i \to \Delta_0$, $\|\alpha_i\|_2 \to 0$, and $\alpha_i(t) \to 0$ almost everywhere $t \in [a, b]$ as $i \to \infty$. For each i, let $u_i \in \Sigma_v(\Delta_i, \alpha_i)$. Passing to a subsequence if necessary, and in view of Lemma 4.4, we may assume $u_i(t) \to v(t)$ almost everywhere $t \in [a, b]$. For the sake of brevity, let $x_i(\cdot) = x_{\alpha_i}^{u_i, \Delta_i}(\cdot), x(\cdot) = x^{v, \Delta_0}(\cdot)$, $(p_i(\cdot), q_i(\cdot)) = M(u_i, P_v(\Delta_i, \alpha_i))$, and $(p(\cdot), q(\cdot)) = M(v, P(\Delta_0))$.

LEMMA 4.5. *As $i \to \infty$, we have*

(a) $x_i(t) \to x(t)$ *uniformly over* $t \in [a, b]$,

(b) $\dot{x}_i(t) \to \dot{x}(t)$ *almost everywhere* $t \in [a, b]$.

Also, subsequences of $\{p_i\}$, $\{q_i\}$ (that are not relabeled) satisfy

(c) $q_i(t) \to q(t)$, $p_i(t) \to p(t)$ *uniformly over* $t \in [a, b]$,

(d) $\dot{q}_i(t) \to \dot{q}(t)$, $\dot{p}_i(t) \to \dot{p}(t)$ *almost everywhere* $t \in [a, b]$.

Proof. By (H5), there exists $r > 0$ so that $\|x_i\|_\infty \leq r$ for all i. Therefore from (H6) and $\dot{x}_i = \phi + \alpha_i$, we have $|\dot{x}_i(t)| \leq \sigma(t)$ for some $\sigma \in L^1[a, b]$. The Dunford–Pettis criterion implies that a subsequence (that is not relabeled) of $\{\dot{x}_i(\cdot)\}_{i=1}^{\infty}$ has a weak limit point $\dot{z}(\cdot)$, whence it immediately follows that $z(t) := c(a) + \int_a^t \dot{z}(s)\,\mathrm{ds}$ satisfies $x_i(t) \to z(t)$ for each $t \in [a, b]$. In fact, the convergence $x_i(t) \to z(t)$ is uniform over $t \in [a, b]$ since $|\dot{x}_i(t)|$ is bounded independent of i by an L^1 function. Then for almost all $t \in [a, b]$, we have $\dot{x}_i(t) = \phi(t, \Delta_i, x_i(t), x_i(t - \Delta_i), u_i(t)) + \alpha_i(t) \longrightarrow \phi(t, \Delta_0, z(t), z(t - \Delta_0), v(t))$ as $i \to \infty$, where we set $z(t) = c(t)$ if $t < a$. In particular, then, $\{\dot{x}_i(\cdot)\}$ converges weakly in $L^1[a, b]$ and almost everywhere on $[a, b]$, and by the uniqueness of weak limits, the limit functions must coincide. Thus $\dot{z}(t) = \phi(t, \Delta_0, z(t), z(t - \Delta_0), v(t))$ almost everywhere $t \in [a, b]$ with $z(t) = c(t)$ for $t \in [a - \Delta_0, a]$. However, $x(\cdot)$ is the unique function satisfying this equation, so we must have $z = x$. Therefore (a) and (b) hold.

Now consider the convergences in (c) and (d). Since $(p_i, q_i) = M(u_i, P_v(\Delta_i, \alpha_i))$, we have

(4.21)

$$\begin{aligned} -p_i(b) &= \nabla_x \ell(\Delta_i, x_i(b)), \quad q_i(b) = 0, \\ -\dot{p}_i(t) &= \phi_1^*(t, \Delta_i, x_i(t), x_i(t-\Delta_i), u_i(t))(p_i(t) + q_i(t+\Delta_i)) \\ &\qquad - L_1^*(t, \Delta_i, x_i(t), x_i(t-\Delta_i), u_i(t)) \quad \text{a.e. } t \in [a,b], \\ -\dot{q}_i(t) &= \phi_2^*(t, \Delta_i, x_i(t), x_i(t-\Delta_i), u_i(t))(p_i(t) + q_i(t+\Delta_i)) \\ &\qquad - L_2^*(t, \Delta_i, x_i(t), x_i(t-\Delta_i), u_i(t)) \quad \text{a.e. } t \in [a,b]. \end{aligned}$$

Similar arguments that were used to prove (a) and (b) can be used here. We merely sketch this : first obtain a priori bounds on $|p_i| + |q_i|$ from (H6), (4.21), and Gronwall's lemma. Second, it follows from (H6) and (4.21) that $|\dot{p}_i(t)| + |\dot{q}_i(t)| \leq \sigma(t)$ for some $\sigma(\cdot) \in L^1[a,b]$. Thus we can deduce the existence of a weak limit point $(\dot{p}(t), \dot{q}(t))$ of a subsequence of $\{(\dot{p}_i(t), \dot{q}_i(t))\}$ from the Dunford–Pettis criterion, and for which $(p_i(t), q_i(t)) \to (p(t), q(t))$ uniformly over $t \in [a,b]$ as $i \to \infty$. Third, we also have from (4.21) and Lemma 4.4 that $(\dot{p}_i(t), \dot{q}_i(t))$ converges almost everywhere $t \in [a,b]$. By the uniqueness of weak limits, the limiting function must be $(\dot{p}(t), \dot{q}(t))$. It is clear from (4.21) that (p, q) satisfies

$$\begin{aligned} -p(b) &= \nabla_x \ell(\Delta_0, x(b)), \quad q(b) = 0, \\ -\dot{p}(t) &= \phi_1^*(t, \Delta_0, x(t), x(t-\Delta_0), v(t))(p(t) + q(t+\Delta_0)) \\ &\qquad - L_1^*(t, \Delta_0, x(t), x(t-\Delta_0), v(t)) \quad \text{a.e. } t \in [a,b], \\ -\dot{q}(t) &= \phi_2^*(t, \Delta_0, x(t), x(t-\Delta_0), v(t))(p(t) + q(t+\Delta_0)) \\ &\qquad - L_2^*(t, \Delta_0, x(t), x(t-\Delta_0), v(t)) \quad \text{a.e. } t \in [a,b]. \end{aligned}$$

This says that (p, q) satisfies the transversality condition and the adjoint equations associated with $(v, P(\Delta_0))$. To finish the proof of (c) and (d), it is only left to show that the maximum condition also holds.

Recall that the pseudo-Hamiltonian is defined by

$$\mathscr{H}(t, \Delta, x, y, u, p) = \langle \phi(t, \Delta, x, y, u), p \rangle - L(t, \Delta, x, y, u).$$

We must show that for almost all $t \in [a,b]$,

$$\max \mathscr{H}(t, \Delta_0, x(t), x(t-\Delta_0), u, p(t) + q(t+\Delta_0)) \tag{4.22}$$

over $u \in U(t)$ is achieved at $u = v(t)$.

Let $t \in [a,b]$ so that the following four conditions hold:

(i) $\alpha_i(t) \to 0$ as $i \to \infty$,
(ii) $u_i(t) \to v(t)$ as $i \to \infty$,
(iii) ϕ and L are continuous in the other variables whenever t is fixed,
(iv) for each i, define $\Phi_i(u)$ by

$$\Phi_i(u) = \mathscr{H}(t, \Delta_i, x_i(t), x_i(t-\Delta_i), u, p_i(t) + q_i(t+\Delta_i)) - |u - v(t)|^2;$$

then max $\Phi_i(u)$ over $u \in U(t)$ is achieved at $u = u_i(t)$.

The t for which (i)–(iv) hold consists of a set of full measure in $[a,b]$. Define $\Phi(u)$ by

$$\Phi(u) = \mathscr{H}(t, \Delta_0, x(t), x(t-\Delta_0), u, p(t) + q(t+\Delta_0)) - |u - v(t)|^2.$$

A consequence of (i), (iii), and the convergences in parts (a) and (c) is that $\Phi_i(u) \to \Phi(u)$ uniformly over $u \in U(t)$. Obviously, from (iv), we have $\Phi_i(u_i(t)) \geq \Phi_i(u)$ for all $u \in U(t)$. Hence letting $i \to \infty$, we deduce from (ii) that

$$\max \Phi(u) \text{ over } u \in U(t) \text{ is achieved at } u = v(t). \tag{4.23}$$

The difference between (4.23) and (4.22) is that the term $|u - v(t)|^2$ appears in (4.23), but not in (4.22). Now a convexity argument similar to that used in the proof of Lemma 4.2 can be applied here to conclude that (4.22) follows from (4.23). (See (4.18) and the succeeding lines.) □

There is an obvious analogue of Lemma 4.5 that will be used in evaluating the lim inf on the right side of Lemma 4.3 (b). Unlike with (a), an element $v \in \Sigma(\Delta_0)$ is not fixed first, but rather will be obtained from (H8). Suppose that we are given $\{(\Delta_i, \alpha_i)\}$ as in (4.13) and $\{u_i\}$ satisfying $u_i \in \Sigma(\Delta_i, \alpha_i)$. We assume that (H8) holds, so there exists $v \in \mathscr{M}[a,b]$ such that an (unrelabeled) subsequence of $\{u_i\}$ satisfies $u_i(t) \to v(t)$ almost everywhere $t \in [a,b]$. We show that this $v(\cdot)$ is optimal for $P(\Delta_0)$.

LEMMA 4.6. *Let $v(\cdot)$ be as the previous paragraph. Then $v(\cdot) \in \Sigma(\Delta_0)$.*

Proof. Let $x_i(\cdot) = x_{\alpha_i}^{u_i,\Delta_i}(\cdot)$ and $x(\cdot) = x^{v,\Delta_0}(\cdot)$. We can show as in the proof of Lemma (4.5)(a) that $x_i(t) \to x(t)$ uniformly over $t \in [a,b]$. It then follows from the basic assumptions that $J(\Delta_i, x_i, u_i) \to J(\Delta_0, x, v)$ as $i \to \infty$. We have

$$\begin{aligned} V(\Delta_0) = V(\Delta_0, 0) &= \lim_{i\to\infty} V(\Delta_i, \alpha_i) \quad \text{(by Step 1)} \\ &= \lim_{i\to\infty} J(\Delta_i, x_i, u_i) \\ &= J(\Delta_0, x, v). \end{aligned}$$

It is now immediate that $v(\cdot) \in \Sigma(\Delta_0)$. □

The analogue of Lemma 4.5 for the small change in data just introduced can now be discerned. Assertions (a)-(d) hold for $u_i \in \Sigma(\Delta_i, \alpha_i)$, $x_i(\cdot) = x_{\alpha_i}^{u_i,\Delta_i}(\cdot)$, $x(\cdot) = x^{v,\Delta_0}(\cdot)$, $(p_i, q_i) = M(u_i, P(\Delta_i, \alpha_i))$, and $(p,q) = M(v, P(\Delta_0))$.

Step 7. We are ready to address the convergence of the integrals in Lemma 4.3. The next two lemmas are given abstract formulation so that they can also be used in §6. They concern weak convergence with deviating arguments.

LEMMA 4.7. *Suppose that $\psi(\cdot) \in L^2[a,b]$, $\{\psi_i(\cdot)\} \subseteq L^2[a,b]$ satisfy $|\psi_i(t)| \leq \gamma(t)$ for some $\gamma(\cdot) \in L^2[a,b]$, and $\psi_i(\cdot) \to \psi(\cdot)$ weakly in $L^2[a,b]$. Let $\{h_i\}_{i=1}^{\infty} \subseteq \mathbb{R}^1$ be such that $h_i \to 0$ as $i \to \infty$. Define $\Psi_i(t) = \psi_i(t - h_i)$ if $t - h_i \in [a,b]$, and let $\Psi_i(t)$ take on any values with absolute value less than $\gamma(t)$ if $t - h_i \notin [a,b]$ (and so that $\Psi_i(\cdot)$ remains measurable). Then $\Psi_i(\cdot) \to \psi(\cdot)$ weakly in $L^2[a,b]$.*

Proof. Let $\epsilon > 0$ and $g \in L^2[a,b]$. There exists continuous f on $[a,b]$ such that $||f - g||_2 < \epsilon$. We have

$$\begin{aligned} \left| \int_a^b \langle g(t), \Psi_i(t) - \psi(t) \rangle \, \mathrm{dt} \right| &\leq ||f - g||_2 \, ||\Psi_i(t) - \psi(t)||_2 + \left| \int_a^b \langle f(t), \Psi_i(t) - \psi(t) \rangle \, \mathrm{dt} \right| \\ &\leq \epsilon(||\gamma||_2 + ||\psi||_2) + \left| \int_a^b \langle f(t), \Psi_i(t) - \psi(t) \rangle \, \mathrm{dt} \right|. \end{aligned} \tag{4.24}$$

Let $a_i = a + |h_i|$ and $b_i = b - |h_i|$. Then for each i we have

$$\begin{aligned}\int_a^b \langle f(t), \Psi_i(t)\rangle \,\mathrm{dt} &= \int_{a_i}^{b_i} \langle f(t), \psi_i(t-h_i)\rangle \,\mathrm{dt} + \int_{[a,b]\setminus[a_i,b_i]} \langle f(t), \Psi_i(t)\rangle \,\mathrm{dt} \\ &= \int_{a_i-h_i}^{b_i-h_i} \langle f(t+h_i) - f(t), \psi_i(t)\rangle \,\mathrm{dt} - \int_{[a,b]\setminus[a_i-h_i, b_i-h_i]} \langle f(t), \psi_i(t)\rangle \,\mathrm{dt} \\ &\quad + \int_{[a,b]\setminus[a_i,b_i]} \langle f(t), \Psi_i(t)\rangle \,\mathrm{dt} + \int_a^b \langle f(t), \psi_i(t)\rangle \,\mathrm{dt}.\end{aligned}$$

The dominated convergence theorem implies that each of the first three integrals on the right-hand side goes to zero as $i \to \infty$. The fourth integral converges to $\int_a^b \langle f(t), \psi(t)\rangle \,\mathrm{dt}$ as $i \to \infty$, since $\psi_i \to \psi$ weakly. We conclude from (4.24) that $\Psi_i \to \psi$ weakly in $L^2[a, b]$. □

LEMMA 4.8. *Suppose that $\psi(\cdot)$, $\{\psi_i(\cdot)\}_{i=1}^{\infty}$, and $\{\Psi_i\}_{i=1}^{\infty}$ are as in Lemma 4.7. Suppose further that functions $g(\cdot)$ and $\{g_i(\cdot)\}_{i=1}^{\infty}$ are given so that $|g_i(t)| \le \sigma(t)$ for some $\sigma(\cdot) \in L^2[a, b]$, and that $g_i(t) \to g(t)$ almost everywhere $t \in [a, b]$ as $i \to \infty$. Then*

$$\int_a^b \langle g_i(t), \Psi_i(t)\rangle \,\mathrm{dt} \to \int_a^b \langle g(t), \psi(t)\rangle \,\mathrm{dt}$$

as $i \to \infty$.

Proof. By Lemma 4.7, we have

$$\int_a^b \langle g(t), \Psi_i(t)\rangle \,\mathrm{dt} \to \int_a^b \langle g(t), \psi(t)\rangle \,\mathrm{dt} \tag{4.25}$$

as $i \to \infty$. By assumption, $|\langle g_i(t) - g(t), \Psi_i(t)\rangle|$ is bounded almost everywhere $t \in [a, b]$ by an $L^1[a, b]$ function and goes to zero almost everywhere $t \in [a, b]$ as $i \to \infty$. Hence the dominated convergence theorem and (4.25) yield

$$\begin{aligned}\int_a^b \langle g_i(t), \Psi_i(t)\rangle \,\mathrm{dt} &= \int_a^b \langle g_i(t) - g(t), \Psi_i(t)\rangle \,\mathrm{dt} + \int_a^b \langle g(t), \Psi_i(t)\rangle \,\mathrm{dt} \\ &\to \int_a^b \langle g(t), \psi(t)\rangle \,\mathrm{dt}\end{aligned}$$

as $i \to \infty$. This proves the lemma. □

COROLLARY 4.9. *Let $\{x_i\}$, $\{q_i\}$, x, and q be as in Lemma 4.5. Then as $i \to \infty$, we have*

$$\int_a^b \langle \dot{q}_i(t), \dot{x}_i(t - \Delta_i)\rangle \,\mathrm{dt} \to \int_a^b \langle \dot{q}(t), \dot{x}(t - \Delta_0)\rangle \,\mathrm{dt}.$$

PROOF: Apply Lemma 4.8 to $\psi(t) = \dot{x}(t - \Delta_0)$, $\psi_i(t) = \dot{x}_i(t - \Delta_0)$, $h_i = \Delta_i - \Delta$, $g(t) = \dot{q}(t)$, and $g_i(t) = \dot{q}_i(t)$. It is routine to check hypotheses, and the conclusion follows directly. □

End of the proof of Theorem 3.1. Recall Lemma 4.3(a). Fix $v \in \Sigma(\Delta_0)$. Let $\{(\Delta_i, \alpha_i)\}_{i=1}^{\infty}$ be a sequence for which $(\Delta_i, \alpha_i) \to (\Delta_0, 0)$ in norm and V_v has a proximal subgradient at (Δ_i, α_i). Let $u_i \in \Sigma_v(\Delta_i, \alpha_i)$. We pass to a subsequence if necessary so that $\alpha_i(t) \to 0$, $u_i(t) \to v(t)$ almost everywhere $t \in [a,b]$, and so that the convergences in Lemma 4.5 are valid. It follows from (H1) that

$$\frac{\partial}{\partial\Delta}\ell(\Delta_i, x_i(b)) \to \frac{\partial}{\partial\Delta}\ell(\Delta_0, x(b)). \tag{4.26}$$

From (H2), (H3), and Lemmas 4.4 and 4.5 (a), (c), we conclude that

$$(\partial/\partial\Delta)\mathscr{H}(t, \Delta_i, x_i(t), x_i(t-\Delta_i), u_i(t), p_i(t) + q_i(t+\Delta_i))$$

converges for almost all $t \in [a,b]$ to

$$(\partial/\partial\Delta)\mathscr{H}(t, \Delta_0, x(t), x(t-\Delta_0), v(t), p(t) + q(t+\Delta_0))$$

as $i \to \infty$. Since each of these functions is bounded by an $L^1[a,b]$ function independent of i, the dominated convergence theorem implies

$$\begin{aligned}\int_a^b \frac{\partial}{\partial\Delta}&\mathscr{H}(t, \Delta_i, x_i(t), x_i(t-\Delta_i), u_i(t), p_i(t) + q_i(t+\Delta_i))\,\mathrm{dt}\\ &\to \int_a^b \frac{\partial}{\partial\Delta}\mathscr{H}(t, \Delta_0, x(t), x(t-\Delta_0), v(t), p(t) + q(t+\Delta_0))\,\mathrm{dt}\end{aligned} \tag{4.27}$$

as $i \to \infty$. Finally, from Corollary 4.9, we have

$$\int_a^b \langle \dot{q}_i(t), x_i(t-\Delta_i)\rangle\,\mathrm{dt} \to \int_a^b \langle \dot{q}(t), x(t-\Delta_0)\rangle\,\mathrm{dt} \tag{4.28}$$

as $i \to \infty$. Now the convergences (4.26), (4.27), and (4.28) hold for an arbitrary $v \in \Sigma(\Delta_0)$. Hence we conclude from Lemma 4.3(a) that

$$\begin{aligned}\limsup_{h\searrow 0} \frac{V(\Delta_0+h) - V(\Delta_0)}{h} \le \inf_{v\in\Sigma(\Delta_0)} \Bigg\{ &\frac{\partial}{\partial\Delta}\ell(\Delta_0, x(b))\\ &- \int_a^b \left[\frac{\partial}{\partial\Delta}\mathscr{H}\Big(t, \Delta_0, x(t), x(t-\Delta_0), v(t), p(t) + q(t+\Delta_0)\Big)\right.\\ &\left. + \ \langle \dot{q}(t), \dot{x}(t-\Delta_0)\rangle\right]\mathrm{dt}\Bigg\},\end{aligned} \tag{4.29}$$

where $x(\cdot) = x^{v,\Delta_0}(\cdot)$ and $(p(\cdot), q(\cdot)) = M(v, P(\Delta_0))$. This finishes the proof of Theorem 3.1 (a). □

The proof of (b) is analogous to that of (a). The only significant change is in Step 2, where we note that

$$\begin{aligned}\limsup_{h\searrow 0} \frac{V(\Delta_0-h) - V(\Delta_0)}{h} &\le \inf_{v\in\Sigma(\Delta_0)} -\frac{V_v(\Delta_0) - V_v(\Delta_0-h)}{h}\\ &\le - \sup_{v\in\Sigma(\Delta_0)} V_v^0((\Delta_0, 0); (1,0)).\end{aligned}$$

The rest of the proof proceeds as before; note that the statement of Theorem 3.1(b) ensues from replacing h above by $-h$. □

End of the proof of Theorem 3.2. To prove Theorem 3.2 (a), it suffices to show that under assumption (H8), the right-hand side of Lemma 4.3 (b) equals the right-hand side in (4.29), since, in view of Theorem 3.1(a), the limit exists and is equal to precisely that quantity.

Let $\{(\Delta_i,\alpha_i)\}_{i=1}^{\infty}$ be such that $(\Delta_i,\alpha_i)\to(\Delta_0,0)$ and V has a proximal subgradient at (Δ_i,α_i). Let $u_i\in\Sigma(\Delta_i,\alpha_i)$, and invoke (H8). We obtain a subsequence (not relabeled) and $v(\cdot)\in\mathscr{M}[a,b]$ so that $u_i(t)\to v(t)$ almost everywhere $t\in[a,b]$. By Lemma 4.6, we have $v(\cdot)\in\Sigma(\Delta_0)$. As pointed out at the end of Step 6, the assertions of Lemma 4.5 are valid with the data introduced here, and thus Corollary 4.9 is also valid. All of this leads directly to the convergences in (4.26), (4.27), and (4.28) for an appropriate subsequence. Therefore, from Lemma 4.3 (b), we conclude that

$$\begin{aligned}(4.30)\quad \liminf_{h\searrow 0}\frac{V(\Delta_0+h)-V(\Delta_0)}{h}\ge \inf_{v\in\Sigma(\Delta_0)}\Bigg\{&\frac{\partial}{\partial\Delta}\ell(\Delta_0,x(b))\\ &-\int_a^b\Bigg[\frac{\partial}{\partial\Delta}\mathscr{H}\Big(t,\Delta_0,x(t),x(t-\Delta_0),v(t),p(t)+q(t+\Delta_0)\Big)\\ &+\langle\dot q(t),\dot x(t-\Delta_0)\rangle\Bigg]\,\mathrm{dt}\Bigg\},\end{aligned}$$

where $x(\cdot)=x^{v,\Delta_0}(\cdot)$ and $(p(\cdot),q(\cdot))=M(v,P(\Delta_0))$. The conjunction of (4.30) with (4.29) finishes the proof of Theorem 3.2 (a).

The proof of (b) requires the same modification as the one made in the proof of Theorem 3.1 (b). □

Proof with $\Delta_0=0$. The only difficulty in letting $\Delta_0=0$ occurs in Step 2. It is not permissible there to have the delay parameter be less than zero. However, we can do the following (the notation is as in Step 2):

$$\begin{aligned}\limsup_{h\searrow 0}\frac{V(h)-V(0)}{h}&\le\limsup_{h\searrow 0}\frac{V_v(h)-V_v(0)}{h}\\ &\le\limsup_{\substack{h\searrow 0\\ \Delta\searrow 0\\ \alpha\to 0}}\frac{V_v(\Delta+h,\alpha)-V_v(\Delta,\alpha)}{h}\\ (4.31)\qquad&\le\limsup_{\Delta\searrow 0}\limsup_{\substack{h\searrow 0\\ \Delta'\to\Delta\\ \alpha\to 0}}\frac{V_v(\Delta'+h,\alpha)-V_v(\Delta',\alpha)}{h}\\ &\le\limsup_{\Delta\searrow 0}V_v^0((\Delta,0);(1,0))\\ &=\limsup_{\Delta\searrow 0}\ \sup\mu,\end{aligned}$$

where the last sup is taken over $(\mu,\beta)\in\mathbb{R}^1\times L^2[a,b]$, $\{(\mu_i,\beta_i)\}_{i=1}^{\infty}\subseteq\mathbb{R}^1\times L^2[a,b]$, $\{(\Delta_i,\alpha_i)\}_{i=1}^{\infty}\subseteq(0,\overline{\Delta})\times L^2[a,b]$ with $(\Delta_i,\alpha_i)\to(\Delta,0)$ in norm, $(\mu_i,\beta_i)\to(\mu,\beta)$ weakly, and (μ_i,β_i) is a proximal subgradient of $V_v(\cdot,\cdot)$ at (Δ_i,α_i) . A standard diagonalization argument can be used to absorb "$\limsup\limits_{\Delta\searrow 0}$" into the set at which the sup is taken over. That is, the right-hand side of (5.7) equals

$$\sup\Big\{\mu:(\mu,\beta),\{(\mu_i,\beta_i)\}_{i=1}^{\infty}\text{ with }(\Delta_i,\alpha_i)\to(0,0)\text{ in norm, }(\mu_i,\beta_i)\to(\mu,\beta)\text{ weakly, and }(\mu_i,\beta_i)\text{ is proximal subgradient of }V_v(\cdot,\cdot)\text{ at }(\Delta_i,\alpha_i)\Big\}.$$

From here, the proof of Theorem 3.1 proceeds, commencing with Step 3 precisely as before, with Δ_0 replaced by 0 throughout.

The proof of the opposite bounds with $\Delta_0 = 0$ contains no surprises. The slight modification in the limits introduced in (4.31) can be carried over to the situation in Theorem 3.2. Then as expected, the rest of the proof proceeds as before. □

5. Lipschitz terminal cost.

In this section, we consider problems in which the terminal cost function ℓ is assumed to be merely Lipschitz in the state variable. In addition to its intrinsic interest, the results here will be applied in §7, where endpoint constraints are introduced and a nondifferentiable penalization function will be needed to treat them. In place of (H1), we consider

(H1′) $\ell : [0, \overline{\Delta}) \times \mathbb{R}^n \longrightarrow \mathbb{R}^1$ is C^1 in the first and locally Lipschitz in the second variable. Furthermore, the set $\{(\Delta, x, \zeta) : \zeta \in \hat{\partial}_x \ell(\Delta, x)\}$ is closed in $[0, \overline{\Delta}) \times \mathbb{R}^n \times \mathbb{R}^n$.

Recall from (2.4) that $\hat{\partial}$ denotes the presubdifferential. The x subscript indicates that this is taken with respect to the x variable while Δ is held fixed. Obviously, (H1) implies (H1′). The closure statement is equivalent to saying the multifunction $(\Delta, x) \to \hat{\partial}_x \ell(\Delta, x)$ is upper semicontinuous on $[0, \overline{\Delta}) \times \mathbb{R}^n$. We point out that (H1′) is satisfied whenever ℓ is C^1 in Δ and convex in x.

The notation of the previous sections is maintained, except that the multiplier sets must incorporate new boundary conditions. For $u \in \Sigma(\Delta)$, $M(u, P(\Delta))$ is the *set* of pairs of absolutely continuous functions (p, q) that satisfy the transversality condition

$$-p(b) \in \hat{\partial}_x \ell(\Delta, x^{u,\Delta}(b)), \quad q(b) = 0, \tag{5.1}$$

and also satisfy the adjoint equations and the maximum condition of the maximum principle. In other words, the difference between what we now call $M(u, P(\Delta))$ and the function with the same designation in the previous sections is that $M(u, P(\Delta))$ includes all functions having the possible boundary conditions (5.1).

The maximum principle is still valid under (H1′). That is, $u \in \Sigma(\Delta)$ implies $M(u, P(\Delta)) \neq \emptyset$. The result of this section is a direct extension of the main results from §3.

THEOREM 5.1. *Assume that* (H1′), (H2)–(H7) *hold and* $\Delta_0 \in (0, \overline{\Delta})$. *We have*

(a)

$$\limsup_{h \searrow 0} \frac{V(\Delta_0 + h) - V(\Delta_0)}{h} \leq \inf_{v \in \Sigma(\Delta_0)} \sup_{(p,q) \in M(v, P(\Delta_0))} E\Big(\Delta_0, x^{v,\Delta_0}, v, (p, q)\Big),$$

and if $\Delta_0 > 0$, *then*

$$\liminf_{h \nearrow 0} \frac{V(\Delta_0 + h) - V(\Delta_0)}{h} \geq \sup_{v \in \Sigma(\Delta_0)} \inf_{(p,q) \in M(v, P(\Delta_0))} E\Big(\Delta_0, x^{v,\Delta_0}, v, (p, q)\Big);$$

(b) *In addition, assume that* (H8) *holds at* Δ_0. *Then*

$$\liminf_{h \searrow 0} \frac{V(\Delta_0 + h) - V(\Delta_0)}{h} \geq \inf_{v \in \Sigma(\Delta_0)} \inf_{(p,q) \in M(v, P(\Delta_0))} E\Big(\Delta_0, x^{v,\Delta_0}, v, (p, q)\Big),$$

and if $\Delta_0 > 0$, *then*

$$\limsup_{h \nearrow 0} \frac{V(\Delta_0 + h) - V(\Delta_0)}{h} \leq \sup_{v \in \Sigma(\Delta_0)} \sup_{(p,q) \in M(v, P(\Delta_0))} E\Big(\Delta_0, x^{v,\Delta_0}, v, (p, q)\Big).$$

(c) *In addition to the assumptions of* (b), *suppose that for each* $v \in \Sigma(\Delta_0)$, $M(v, P(\Delta_0))$ *consists of a single element. Then for all* $\Delta_0 \in [0, \overline{\Delta})$, $\lim_{h\searrow 0} (V(\Delta_0 + h) - V(\Delta_0)/h)$ *exists, and for all* $\Delta_0 \in (0, \overline{\Delta})$, $\lim_{h\nearrow 0}(V(\Delta_0 + h) - V(\Delta_0)/h)$ *exists, and are equal to the respective quantities as in Theorem* 3.2.

(d) *In addition to the assumptions of* (c), *suppose that* $\Sigma(\Delta_0)$ *consists of the single element* v. *Then, if* $\Delta_0 > 0$, *the two-sided derivative* $(d/d\Delta)V(\Delta_0)$ *exists and equals* $E(\Delta_0, x^{v,\Delta_0}, v, (p, q))$, *where* $M(v, P(\Delta_0)) = \{(p, q)\}$.

PROOF: This is nearly identical to the proof given in §4. First, note that (H1′) in place of (H1) does not affect the proof that V is locally Lipschitz in (Δ, α). The only significant modification needed in §4 occurs in the proof of Lemma 4.5, where we must now pass to another subsequence of $\{p_i\}$ if necessary to assure that $-p_i(b)$ converges to $-p(b)$. That the transversality condition (5.1) holds in the limit follows from (H1′). Since $M(v, P(\Delta))$ may contain more than one element, the lim sup on the right-hand side in Lemma 4.3(a) may no longer be an actual limit, and this is reflected in the theorem by taking the sup over $M(v, P(\Delta_0))$. □

6. The lower bound with Hamiltonian multipliers.

In this section, we use a Hamiltonian approach to obtain a lower bound for the right lower Dini derivative of $V(\Delta)$. The idea here, conceptually disjoint from assuming (H8), is to analyze the limiting behavior of trajectories and multipliers that satisfy a Hamiltonian inclusion. The optimal controls will play no major role, and for this reason we introduce the notation

$$\tilde{\Sigma}(\Delta) := \{x^{v,\Delta}(\cdot) : v \in \Sigma(\Delta)\},$$

$$\tilde{\Sigma}(\Delta, \alpha) := \{x^{v,\Delta}(\cdot) : v \in \Sigma(\Delta, \alpha)\}.$$

First, some preliminary definitions and remarks. Recall that the pseudo-Hamiltonian $\mathscr{H}$ is defined by

$$\mathscr{H}(t, \Delta, x, y, u, p) = \langle p, \phi(t, \Delta, x, y, u)\rangle - L(t, \Delta, x, y, u).$$

We now define the (true) Hamiltonian H by

$$H : [a, b] \times [0, \overline{\Delta}) \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^1,$$

$$H(t, \Delta, x, y, p) = \max\{\mathscr{H}(t, \Delta, x, y, u, p) : u \in U(t)\}.$$

For $\alpha(\cdot) \in L^2[a, b]$, the Hamiltonian associated with $P(\Delta, \alpha)$ is denoted by H^α. Evidently, $H^\alpha = H + \langle p, \alpha(t)\rangle$. The generalized gradient of H with respect to (Δ, x, y, p) (always written as ∂H) can be calculated as follows (see [C1]):

$$(6.1)\quad \partial H(t, \Delta, x, y, p) = \operatorname{cl}\operatorname{co}\left\{\left(\frac{\partial}{\partial\Delta}\mathscr{H}(t, \Delta, x, y, u, p),\right.\right.$$
$$\left.\frac{\partial}{\partial x}\mathscr{H}^*(t, \Delta, x, y, u, p),\ \frac{\partial}{\partial y}\mathscr{H}^*(t, \Delta, x, y, u, p),\ \phi(t, \Delta, x, y, u)\right) :$$
$$\left. u \in U(t) \text{ such that } H(t, \Delta, x, y, p) = \mathscr{H}(t, \Delta, x, y, u, p)\right\}.$$

Throughout this section, we will assume (H1′) rather than (H1). In the definition of the (Hamiltonian) multipliers, we add an extra component w to what is usually done so as to manage the explicit Δ-dependence. Let $x(\cdot) \in \tilde{\Sigma}(\Delta)$. Then the set of Hamiltonian multipliers $\tilde{M}(x, P(\Delta))$ is defined by

$$\begin{aligned}\tilde{M}(x,P(\Delta)) = \Big\{ &(w,p,q) \in AC_{2n+1}[a,b] : \\ &w(a) = 0, \quad -p(b) \in \hat{\partial}_x \ell(\Delta, x(b)), q(b) = 0, \quad \text{and} \\ &(-\dot{w}(t), -\dot{p}(t), -\dot{q}(t), \dot{x}(t)) \in \partial H(t, \Delta, x(t), x(t-\Delta), p(t) + q(t+\Delta)) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{a.e. } t \in [a,b] \Big\}.\end{aligned}$$

Whenever $t + \Delta \geq b$, $q(t+\Delta)$ is set equal to zero. For a problem $P(\Delta, \alpha)$, and with $x(\cdot) \in \tilde{\Sigma}(\Delta, \alpha)$, then $\tilde{M}(x, P(\Delta, \alpha))$ is defined similarly with H replaced by H^α.

Suppose that $v \in \Sigma(\Delta)$, and let $x(\cdot) = x^{v,\Delta}(\cdot)$ and $(p,q) \in M(v, P(\Delta))$. Define $w : [a,b] \to \mathbb{R}^1$ by

$$w(t) = -\int_a^t \frac{\partial}{\partial \Delta} \mathscr{H}(s, \Delta, x(s), x(s-\Delta), v(s), p(s) + q(s+\Delta))\, \mathrm{ds}. \tag{6.2}$$

It follows from the maximum principle and (6.1) that $(w,p,q) \in \tilde{M}(x, P(\Delta))$. In general, $\tilde{M}(x, P(\Delta))$ will contain other elements, which is true even for problems without delay; see Clarke [C2]. It can be shown, however, that if H is C^1 in (Δ, x, y), then $\tilde{M}(x, P(\Delta)) = \{(w,p,q) : (p,q) \in M(v, P(\Delta)),\ w(\cdot) \text{ as in (6.2)}\}$. Indeed, from [C1, Thm. 2.8.2] it follows that H is regular (see [C1, p. 39]), and consequently $\partial H \subseteq \partial_{(\Delta,x,y)} H \times \partial_p H$, where the subscripts on ∂ indicate a partial generalized gradient (see [C1, Prop. 2.3.15]). If H is C^1 in (Δ, x, y), then $\partial_{(\Delta,x,y)} H$ is the ordinary gradient. Thus, in light of (6.1), $(w,p,q) \in \tilde{M}(x, P(\Delta))$ implies that $(p,q) \in M(v, P(\Delta))$ and $w(\cdot)$ is defined by (6.2).

A necessary condition for $x(\cdot) \in \tilde{\Sigma}(\Delta)$ is that $\tilde{M}(x, P(\Delta)) \neq \emptyset$. This necessary optimality condition is applicable under weaker smoothness assumptions than we have hypothesized. In brief, the C^1 regularity in the state variables can be relaxed to mere Lipschitz with the Lipschitz constants satisfying an analogue of (H6). See Clarke and Watkins [CW].

The advantage Hamiltonian multipliers wield over Pontryagin multipliers is that $\tilde{M}(x, P(\Delta))$ has more robust closure properties (see Lemma 6.3 below). An alternative approach, that we do not develop, would be to deal with measure-relaxed problems in the sense of Warga [W1]. In any case, the convergence of the integral term in (3.2) remains problematic. We are only capable of overcoming this obstruction to our proof technique by invoking one of the following hypotheses:

(H9) For almost all $t \in [a,b]$, $(\partial/\partial y) H(t, \Delta, x, y, p)$ exists and is continuous in (Δ, x, y, p).

(H10) For almost all $t \in [a,b]$, $(\partial/\partial p) H(t, \Delta, x, y, p)$ exists and is continuous in (Δ, x, y, p).

Unlike (H8), (H9) or (H10) can often be verified directly and simply. For example, if the control is separated from y, then (H9) holds. That is, suppose that $\mathscr{H}$ has the form $\mathscr{H}(t, \Delta, x, y, u, p) = f_1(t, \Delta, x, y, p) + f_2(t, \Delta, x, u, p)$ where f_1, f_2 are measurable in t, smooth in (Δ, x, y, p), and f_2 is continuous in u. Then by using (6.1), we can easily verify (H9). Mayer problems with linear dynamics and the example considered in §8C exhibit this structure. An instance in which (H10) holds is when ϕ is linear in u and $L(t, \Delta, x, y, \cdot)$ is strictly convex.

The functional $\tilde{E}$, which is analogous to E of (3.1), is defined by

$$\tilde{E} : [0,\overline{\Delta}) \times AC_n[a,b] \times \mathbb{R}^1 \times AC_n[a,b] \longrightarrow \mathbb{R}^1,$$

$$\tilde{E}(\Delta, x(\cdot), w, q(\cdot)) = \frac{\partial}{\partial \Delta}\ell(\Delta, x(b)) + w + \int_a^b \langle -\dot{q}(t), \dot{x}(t-\Delta)\rangle \, \mathrm{dt}.$$

The main result of this section follows.

THEOREM 6.1. *Suppose that* (H1′), (H2)-(H7) *hold. Assume further that either* (H9) *or* (H10) *holds. Then for* $\Delta_0 \in [0,\overline{\Delta})$, *we have*

(a) $$\liminf_{h\searrow 0} \frac{V(\Delta_0+h)-V(\Delta_0)}{h} \geq \inf_{x\in\tilde{\Sigma}(\Delta_0)} \inf_{(w,p,q)\in\tilde{M}(x,P(\Delta_0))} \tilde{E}(\Delta_0, x, w(b), q);$$

(b) *If* $\Delta_0 > 0$, *then we have*

$$\limsup_{h\nearrow 0} \frac{V(\Delta_0+h)-V(\Delta_0)}{h} \leq \sup_{x\in\tilde{\Sigma}(\Delta_0)} \sup_{(w,p,q)\in\tilde{M}(x,P(\Delta_0))} \tilde{E}(\Delta_0, x, w(b), q).$$

Proof. We give the details for (a) and with $\Delta_0 \in (0,\overline{\Delta})$; part (b) and the case $\Delta_0 = 0$ are handled as they were in §4. We commence as in §4 through Lemma 4.3, where we now are only interested in the lower bound in Lemma 4.3 (b). Suppose that $\{(\Delta_i,\alpha_i)\}_{i=1}^{\infty} \subseteq [0,\overline{\Delta}) \times L^2[a,b]$ satisfies $(\Delta_i,\alpha_i) \to (\Delta_0, 0)$ as $i\to\infty$, and that V has a proximal subgradient at (Δ_i,α_i). Let $u_i \in \Sigma(\Delta_i,\alpha_i)$ and $(p_i,q_i) \in M(u_i, P(\Delta_i,\alpha_i))$, and set $x_i(\cdot) = x_{\alpha_i}^{u_i,\Delta_i}(\cdot)$. Define $w_i : [a,b] \to \mathbb{R}^1$ by

$$w_i(t) = -\int_a^t \frac{\partial}{\partial\Delta}\mathscr{H}(s,\Delta_i,x_i(s),x_i(s-\Delta_i),u_i(s),p_i(s)+q_i(s+\Delta_i))\,\mathrm{ds}.$$

Then we have $(w_i,p_i,q_i) \in \tilde{M}(x_i, P(\Delta_i,\alpha_i))$ for each i. As used in the proof of Lemma 4.5, we can obtain a priori bounds on $|x_i(t)|$, $|p_i(t)|$, and $|q_i(t)|$ independent of i and $t\in[a,b]$. Note from (H6) that $|(\dot{w}_i(t),\dot{p}_i(t),\dot{q}_i(t),\dot{x}_i(t))| \leq \gamma(t)$ for some $\gamma(\cdot) \in L^1[a,b]$. Then the Dunford–Pettis criterion provides a subsequence of $\{(w_i,p_i,q_i,x_i)\}$ (which is not relabeled) and absolutely continuous functions w,p,q, and x so that $(w_i(t),p_i(t),q_i(t),x_i(t)) \to (w(t),p(t),q(t),x(t))$ uniformly over $t\in[a,b]$, and $(\dot{w}_i,\dot{p}_i,\dot{q}_i,\dot{x}_i) \to (\dot{w},\dot{p},\dot{q},\dot{x})$ weakly in $L^1_{3n+1}[a,b]$ as $i\to\infty$.

LEMMA 6.2. *Let* $x(\cdot)$ *be as in the last paragraph. Then* $x(\cdot) \in \tilde{\Sigma}(\Delta_0)$.

Proof. The following is adapted from a standard existence theory argument. Let $F : [a,b]\times[0,\overline{\Delta})\times\mathbb{R}^n\times\mathbb{R}^n \rightrightarrows \mathbb{R}^n\times\mathbb{R}^1$ be the multifunction defined by $F(t,\Delta,x,y) = \{(\phi(t,\Delta,x,y,u), L(t,\Delta,x,y,u)+\delta) : u\in U(t),\ \delta\geq 0\}$. The values of F are closed and convex by (H7). Also, F is measurable in t and continuous in the other variables (see [C1, §3.1]). Define

$$z_i(t) = \int_a^t L(s,\Delta_i,x_i(s),x_i(s-\Delta_i),u_i(s))\,\mathrm{ds}.$$

Then

$$(\dot{x}_i(s)-\alpha_i(s), \dot{z}_i(s)) \in F(s,\Delta_i,x_i(s),x_i(s-\Delta_i)) \quad \text{a.e. } s\in[a,b].$$

Note that $\dot{x}_i(\cdot)-\alpha_i(\cdot) \to \dot{x}(\cdot)$ weakly in $L^1[a,b]$, and passing to another subsequence if necessary, there exists $z(\cdot)\in AC[a,b]$ so that $z_i(t)\to z(t)$ uniformly over $t\in[a,b]$, and $\dot{z}_i(\cdot)\to\dot{z}(\cdot)$ weakly in $L^1[a,b]$. Furthermore, the convex analysis argument used in the proof of [C1, Thm. 3.1.7] can be applied here with only minor modifications (to incorporate the delay Δ and delay component y), and this yields that

$$(\dot{x}(t),\dot{z}(t)) \in F(t,\Delta_0,x(t),x(t-\Delta_0)) \quad \text{a.e. } t\in[a,b].$$

Consequently, (x, z) can be represented from the Filippov lemma as the solution of

$$\dot{x}(t) = \phi(t, \Delta_0, x(t), x(t - \Delta_0), v(t)), \tag{6.3}$$

$$\dot{z}(t) = L(t, \Delta_0, x(t), x(t - \Delta_0), v(t)) + \delta(t), \tag{6.4}$$

where $v(\cdot) \in \mathscr{M}[a, b]$ satisfies $v(t) \in U(t)$ almost everywhere $t \in [a, b]$ and $\delta(\cdot)$ is measurable with $\delta(t) \geq 0$ almost everywhere $t \in [a, b]$. We have

$$\begin{aligned} V(\Delta_0) &= \lim_{i \to \infty} V(\Delta_i, \alpha_i) \\ &= \lim_{i \to \infty} \ell(\Delta_i, x_i(b)) + z_i(b) \qquad (\text{since } u_i \in \Sigma(\Delta_i, \alpha_i)) \\ &= \ell(\Delta_0, x(b)) + z(b) \end{aligned} \tag{6.5}$$

Moreover, it follows from (6.4) that

$$\begin{aligned} z(b) &\geq \int_a^b L(t, \Delta_0, x(t), x(t - \Delta_0), v(t))\, \mathrm{d}t \\ &\geq V(\Delta_0) - \ell(\Delta_0, x(b)) \qquad (\text{by (6.3) and definition of } V) \\ &= z(b). \qquad (\text{by (6.5)}). \end{aligned}$$

The conclusion is that $V(\Delta_0) = J(\Delta_0, x, v)$, and since $x(\cdot) = x^{v, \Delta_0}(\cdot)$ by (6.3), we deduce that $x(\cdot) \in \tilde{\Sigma}(\Delta_0)$. □

LEMMA 6.3. *Let (w, p, q, x) be as above. Then $(w, p, q) \in \tilde{M}(x, P(\Delta_0))$.*

Proof. For each i, the transversality condition

$$w_i(a) = 0, \quad -p_i(b) \in \hat{\partial}_x \ell(\Delta_i, x_i(b)), \quad q_i(b) = 0$$

is satisfied. Now $(w_i(a), p_i(b), q_i(b), x_i(b), \Delta_i) \to (w(a), p(b), q(b), x(b), \Delta_0)$ as $i \to \infty$. Hence by (H1′), we have

$$w(a) = 0, \quad -p(b) \in \hat{\partial}_x \ell(\Delta_0, x(b)), \quad q(b) = 0. \tag{6.6}$$

Also, for each i and almost all $t \in [a, b]$, the Hamiltonian inclusion

$$(-\dot{w}_i(t), -\dot{p}_i(t), -\dot{q}_i(t), \dot{x}_i(t)) \in \partial H^{\alpha_i}(t, \Delta_i, x_i(t), x_i(t - \Delta_i), p_i(t) + q_i(t + \Delta_i)) \tag{6.7}$$

holds. Recall that $H^\alpha = H + \langle p, \alpha(t) \rangle$. Using (6.1), we easily see that $\partial H^\alpha \subseteq \partial H = (0, 0, 0, \alpha(t))$, where $(0, 0, 0, \alpha(t)) \in \mathbb{R}^1 \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n$. Without loss of generality, $\alpha_i(t) \to 0$ as $i \to \infty$ almost everywhere $t \in [a, b]$. The modification in the proof of [C1, Thm. 3.1.7] to delay problems, which was also used in the proof of Lemma 6.2, implies that the Hamiltonian inclusion (6.7) is preserved in the limit as $i \to \infty$. That is, we have

$$(-\dot{w}(t), -\dot{p}(t), -\dot{q}(t), \dot{x}(t)) \in \partial H(t, \Delta_0, x(t), x(t - \Delta_0), p(t) + q(t + \Delta_0)) \quad \text{a.e. } t \in [a, b]. \tag{6.8}$$

By definition, (6.6) and (6.8) say that $(w, p, q) \in \tilde{M}(x, P(\Delta_0))$. □

We now come to the part of the proof where an extra hypothesis is required.

LEMMA 6.4. *Suppose that either* (H9) *or* (H10) *holds. Then*

$$\int_a^b \langle \dot{q}_i(t), \dot{x}_i(t-\Delta_i\rangle \, \mathrm{dt} \to \int_a^b \langle \dot{q}(t), \dot{x}(t-\Delta_0)\rangle \, \mathrm{dt} \tag{6.9}$$

as $i \to \infty$.

Proof. Recall that $(\dot{q}_i(\cdot), \dot{x}_i(\cdot)) \to (\dot{q}(\cdot), \dot{x}(\cdot))$ weakly in $L^2_{2n}[a,b]$ and $(p_i(t), q_i(t), x_i(t)) \to (p(t), q(t), x(t))$ uniformly over $t \in [a,b]$. It was mentioned in the preliminaries of this section that H and H^α are regular, and so

$$\partial H^\alpha \subseteq \partial_{(\Delta,x)} H \times \partial_y H \times (\alpha(t) + \partial_p H). \tag{6.10}$$

Suppose that (H9) holds, and thus $\partial_y H = \{\partial/\partial y H\}$ almost everywhere $t \in [a,b]$. Since $(w_i, p_i, q_i) \in \tilde{M}(x_i, P(\Delta_i, \alpha_i))$, we have by (6.10) that

$$-\dot{q}_i(t) = \partial/\partial y H(t, \Delta_i, x_i(t), x_i(t-\Delta_i), p_i(t) + q_i(t+\Delta_i)) \quad \text{a.e. } t \in [a,b]. \tag{6.11}$$

Since $(w,p,q) \in \tilde{M}(x, P(\Delta_0))$ (Lemma 6.3), we also have that

$$-\dot{q}(t) = \frac{\partial}{\partial y} H(t, \Delta_0, x(t), x(t-\Delta_0), p(t) + q(t+\Delta_0)) \quad \text{a.e. } t \in [a,b]. \tag{6.12}$$

By invoking (H9), we have that the right-hand side of (6.11) converges almost everywhere $t \in [a,b]$ to the right-hand side (6.12), wherefore $\dot{q}_i(t) \to \dot{q}(t)$ almost everywhere $t \in [a,b]$ as $i \to \infty$. Now the assumptions of Lemma 4.8 for the data $g_i = \dot{q}_i$, $g = \dot{q}$, $\psi_i(t) = \dot{x}_i(t-\Delta_0)$, $\psi(t) = \dot{x}(t-\Delta_0)$, $h_i = \Delta_i - \Delta_0$ can be easily verified. The conclusion of Lemma 4.8 with this data is (6.9).

Now suppose that (H10) holds, and thus $\partial_p H = \{\partial/\partial p H\}$ almost everywhere $t \in [a,b]$. Since $(w_i, p_i, q_i) \in M(x_i, P(\Delta_i, \alpha_i))$, we have by (6.10) that

$$\dot{x}_i(t) = \frac{\partial}{\partial p} H(t, \Delta_i, x_i(t), x_i(t-\Delta_i), p_i(t) + q_i(t+\Delta_i)) + \alpha_i(t) \quad \text{a.e. } t \in [a,b]. \tag{6.13}$$

From (6.10) (with $\alpha = 0$) and Lemma 6.3, it follows that

$$\dot{x}(t) = \frac{\partial}{\partial p} H(t, \Delta_0, x(t), x(t-\Delta_0), p(t) + q(t+\Delta_0)) \quad \text{a.e. } t \in [a,b]. \tag{6.14}$$

Since $\alpha_i(t) \to 0$ almost everywhere $t \in [a,b]$, (H10) implies that for almost all $t \in [a,b]$, the right-hand side of (6.13) approaches that in (6.14). Therefore $\dot{x}_i(t) \to \dot{x}(t)$ almost everywhere $t \in [a,b]$ as $i \to \infty$. Now we note that the assumptions of Lemma 4.8 are satisfied for the data $g_i = \dot{x}_i$, $g = \dot{x}$, $\psi_i(t) = \dot{q}_i(t+\Delta_0)$, $\psi(t) = \dot{q}(t+\Delta_0)$, $h_i = \Delta_i - \Delta_0$. The conclusion of Lemma 4.8 again implies that (6.9) holds. □

End of proof. Recall the lower bound in Lemma 4.3 (b). Let us set

$$\mu_i := \tilde{E}(\Delta_i, x_i, w_i(b), q_i).$$

We have shown that there exists $x(\cdot) \in \tilde{\Sigma}(\Delta_0)$ (see Lemma 6.2) and $(w,p,q) \in \tilde{M}(x, P(\Delta_0))$ (Lemma 6.3) such that $\mu_i \to \mu$ (Lemma 6.4), where

$$\mu := \tilde{E}(\Delta_0, x, w(b), q).$$

Now the sequence $\{(\Delta_i, \alpha_i)\}_{i=1}^\infty$ was chosen arbitrarily among those in which the lim inf is taken in Lemma 4.3 (b). Hence we conclude from Lemma 4.3 (b) that

$$\liminf_{h\to 0} \frac{V(\Delta_0 + h) - V(\Delta_0)}{h} \geq \inf_{x \in \tilde{\Sigma}(\Delta_0)} \; \inf_{(w,p,q) \in \tilde{M}(x, P(\Delta_0))} \tilde{E}(\Delta_0, x, w(b), q).$$

This was the assertion of the theorem. □

Remark. Assumptions (H9) and (H10) are stronger than what was required in the proof. The same conclusion of Theorem 6.1 can be drawn if we assume in place of (H9) that for fixed $\Delta_0 \in [0, \overline{\Delta})$, $(\partial/\partial y)H(t, \Delta_0, x, y, p)$ exists and is the only element in the set limit of $\partial_y H(t, \Delta_i, x_i, y_i, p_i)$ as $(\Delta_i, x_i, y_i, p_i) \to (\Delta_0, x, y, p)$. This must hold for almost all $t \in [a, b]$, and for $(x, y, p) \in \mathbb{R}^{3n}$. We can make a similar replacement for (H10).

In many situations, it may be known that the multiplier set associated with an optimal trajectory consists of only one element. This is true in two examples considered in §8. We give this property a precise statement. Consider the following.

(H11) Fix $\Delta_0 \in [0, \overline{\Delta})$. For each $x \in \tilde{\Sigma}(\Delta_0)$, $\tilde{M}(x, P(\Delta_0))$ contains only one element.

An immediate corollary to Theorems 3.1 and 6.1 is the following.

COROLLARY 6.5. *Let* $\Delta_0 \in [0, \overline{\Delta})$ *and suppose* (H1′), (H2)–(H7), (H11), *and either* (H9) *or* (H10) *holds. Then we have the following:*

(a) *The right derivative of* V *at* Δ_0 *exists, and*

$$\lim_{h \searrow 0} \frac{V(\Delta_0 + h) - V(\Delta_0)}{h} = \inf_{x \in \tilde{\Sigma}(\Delta_0)} \tilde{E}(\Delta_0, x, w(b), q),$$

where $\tilde{M}(x, P(\Delta_0)) = \{(w, p, q)\}$;

(b) *If* $\Delta_0 > 0$, *then the left derivative of* V *at* Δ_0 *exists, and*

$$\lim_{h \nearrow 0} \frac{V(\Delta_0 + h) - V(\Delta_0)}{h} = \sup_{x \in \tilde{\Sigma}(\Delta_0)} \tilde{E}(\Delta_0, x, w(b), q),$$

where $\tilde{M}(x, P(\Delta_0)) = \{(w, p, q)\}$;

(c) *In addition, suppose* $\tilde{\Sigma}(\Delta_0)$ *consists of a single element* x. *Let* $\{(w, p, q)\} = \tilde{M}(x, P(\Delta_0))$. *Then if* $\Delta_0 > 0$, *the derivative of* V *at* Δ_0 *exists and equals* $\tilde{E}(\Delta, x, w(b), q)$.

Remark. If (H11) is satisfied and $v \in \Sigma(\Delta_0)$, then $M(v, P(\Delta_0))$ also consists of only one element, say $M(v, P(\Delta_0)) = \{(p, q)\}$. Let $w(\cdot)$ be defined as in (6.2). We have that $\tilde{M}(x^{v, \Delta_0}, P(\Delta_0)) = \{(w, p, q)\}$. Note then that the conclusion of Corollary 6.5 is identical with that of Theorem 5.1 (c), (d), since in this case the values of E and $\tilde{E}$ coincide.

7. Endpoint constraints.

The previous theory is developed in this section to incorporate an endpoint constraint. The basic data is augmented by a closed, nonempty subset C of $\mathbb{R}^n$. It is convenient at this point to not allow ℓ to depend explicitly on Δ. In a concluding remark, it is noted under what assumptions Δ-dependence can be treated.

The basic problem we now consider is

$$\min J(\Delta, x, u) \quad \text{over} \quad x(\cdot) \in AC[a,b], \quad u(\cdot) \in \mathscr{M}[a,b],$$

satisfying

$$\begin{aligned} &\dot{x}(t) = \phi(t, \Delta, x(t), x(t-\Delta), u(t)) \quad \text{a.e. } t \in [a,b], \\ &u(t) \in U(t) \quad \text{a.e. } t \in [a,b], \\ &x(t) = c(t) \quad \text{for} \quad t \in [a-\Delta, a], \\ &x(b) \in C. \end{aligned} \tag{7.1}$$

The notation of this section will closely resemble that which was previously used. Problem (7.1) is denoted by $P_C(\Delta)$. The set of optimal solutions, the set of associated optimal trajectories, and the optimal value of $P_C(\Delta)$ are written as $\Sigma_C(\Delta)$, $\tilde{\Sigma}_C(\Delta)$, and $V_C(\Delta)$, respectively. Perturbations of the dynamics no longer require notation, but rather endpoint perturbations play a significant role. For $\xi \in \mathbb{R}^n$, $P_{C+\xi}(\Delta)$ (respectively, $P_{\{\xi\}}(\Delta)$) denotes problem (7.1) with C replaced by $C+\xi$ (respectively, $\{\xi\}$). Similar definitions are made for $\Sigma_{C+\xi}(\Delta)$, $\tilde{\Sigma}_{C+\xi}(\Delta)$, and $V_{C+\xi}(\Delta)$. If no feasible solution exists for $P_C(\Delta)$, then the convention $V_C(\Delta) = +\infty$ is adopted. If $V_C(\Delta) < +\infty$, then standard compactness arguments show that $\Sigma_C(\Delta) \neq \emptyset$.

Significant alterations occur in defining the multipliers. Recall that the distance function to a closed set $D \subseteq \mathbb{R}^n$ is defined by

$$d_D(\xi) = \inf\{|\xi - \xi'| : \xi' \in D\},$$

and is globally Lipschitz of order 1. For $\xi \in D$, the prenormal cone $\hat{N}_D(\xi)$ of D at ξ is given by

$$\hat{N}_D(\xi) := \text{cl} \bigcup_{k>0} k\, \hat{\partial} d_D(\xi).$$

Hamiltonians for $P_C(\Delta)$ are indexed by a nonnegative real parameter. For $\lambda \geq 0$, $H^\lambda : [a,b] \times [0, \overline{\Delta}) \times \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}^1$ is defined as

$$H^\lambda(t, \Delta, x, y, p) := \sup\left\{\langle p, \phi(t, \Delta, x, y, u)\rangle - \lambda L(t, \Delta, x, y, u) : u \in U(t)\right\}.$$

Suppose that $u(\cdot) \in \Sigma_C(\Delta)$ and $x(\cdot) = x^{u,\Delta}(\cdot) \in \tilde{\Sigma}_C(\Delta)$. The multiplier sets $M^\lambda(u, P_C(\Delta))$ and $\tilde{M}^\lambda(x, P_C(\Delta))$ of index $\lambda \geq 0$ associated with u and x are defined by

$$\begin{aligned} M^\lambda(u, P_C(\Delta)) := \Bigg\{ &(p,q) \in AC_{2n}[a,b] : \\ &-p(b) \in \lambda\, \hat{\partial}\ell(x(b)) + \hat{N}_C(x(b)), \quad q(b) = 0, \\ &-\dot{p}(t) = \phi_1^*(t, \Delta, x(t), x(t-\Delta), u(t))(p(t) + q(t+\Delta)) \end{aligned}$$

$$- \lambda L_1^*(t,\Delta,x(t),x(t-\Delta),u(t)) \quad \text{a.e. } t \in [a,b],$$

$$\begin{aligned} -\dot{q}(t) = \phi_2^*(t,\Delta,x(t),x(t-\Delta),u(t))(p(t)+q(t+\Delta)) \\ - \lambda L_2^*(t,\Delta,x(t),x(t-\Delta),u(t)) \quad \text{a.e. } t \in [a,b], \end{aligned}$$

$$\left.\begin{aligned} &\max \left\{\langle p(t)+q(t+\Delta), \phi(t,\Delta,x(t),x(t-\Delta),u)\rangle - \lambda L(t,\Delta,x(t),x(t-\Delta),u)\right\} \\ &\qquad \text{over } u \in U(t) \text{ occurs at } u = u(t) \quad \text{a.e. } t \in [a,b] \end{aligned}\right\},$$

$$\begin{aligned} \tilde{M}^\lambda(x, P_C(\Delta)) := \Big\{ &(w,p,q) \in AC_{2n+1}[a,b] : \\ &w(a) = 0, \quad -p(b) \in \lambda\,\hat{\partial}\ell(x(b)) + \hat{N}_C(x(b)), \quad q(b) = 0 \\ &(-\dot{w}(t), -\dot{p}(t), -\dot{q}(t), \dot{x}(t)) \in \partial H^\lambda(t,\Delta,x(t),x(t-\Delta),p(t)+q(t+\Delta)) \\ &\qquad \text{a.e. } t \in [a,b] \Big\}. \end{aligned}$$

In the latter definition, the subgradient ∂H^λ is taken with respect to the (Δ, x, y, p) variables.

Suppose that $x(\cdot) \in \tilde{\Sigma}_C(\Delta)$. It is always the case that $p \equiv 0$ (which implies $q \equiv 0$) is a component of an element of $\tilde{M}^0(x, P_C(\Delta))$. The trajectory $x(\cdot)$ is called normal if $\tilde{M}^0(x, P_C(\Delta))$ contains only elements in which the $p(\cdot)$ trajectory is identically equal to zero. The problem $P_C(\Delta)$ is called normal if each $x(\cdot) \in \tilde{\Sigma}_C(\Delta)$ is normal.

Note that since we are assuming ℓ is independent of Δ, assumption (H1$'$) of §5 reduces to

(H1$''$) $\ell : \mathbb{R}^n \to \mathbb{R}^1$ is local Lipschitz.

The main theorem of this section follows.

THEOREM 7.1. *Suppose* (H1$''$), (H2)–(H7) *hold, and let* $\Delta_0 \in [0, \overline{\Delta})$. *Assume that* $V_C(\Delta_0)$ *is finite and* $P_C(\Delta_0)$ *is normal. Then*

(a)

$$\limsup_{h \searrow 0} \frac{V_C(\Delta_0 + h) - V_C(\Delta_0)}{h} \le \inf_{v(\cdot) \in \Sigma_C(\Delta_0)} \; \sup_{(p,q) \in M^1(v, P_C(\Delta_0))} E(\Delta_0, x^{v,\Delta_0}, v, (p,q)),$$

and if $\Delta_0 > 0$, *then*

$$\liminf_{h \nearrow 0} \frac{V_C(\Delta_0 + h) - V_C(\Delta_0)}{h} \ge \sup_{v \in \Sigma_C(\Delta_0)} \; \inf_{(p,q) \in M^1(v, P_C(\Delta_0))} E(\Delta_0, x^{v,\Delta_0}, v, (p,q));$$

(b) *In addition to the basic hypotheses, assume that* (H8) *holds at* Δ_0 *with* Σ *replaced by* Σ_C. *Then*

$$\liminf_{h \searrow 0} \frac{V_C(\Delta_0 + h) - V_C(\Delta_0)}{h} \ge \inf_{v(\cdot) \in \Sigma_C(\Delta_0)} \; \inf_{(p,q) \in M^1(v, P_C(\Delta_0))} E(\Delta_0, x^{v,\Delta_0}, v, (p,q)),$$

and if $\Delta_0 > 0$, *then*

$$\limsup_{h \nearrow 0} \frac{V_C(\Delta_0 + h) - V_C(\Delta_0)}{h} \le \sup_{v(\cdot) \in \Sigma_C(\Delta_0)} \; \sup_{(p,q) \in M^1(v, P_C(\Delta_0))} E(\Delta_0, x^{v,\Delta_0}, v, (p,q));$$

(c) *In addition to the basic hypotheses of* (a), *assume that* (H9) *or* (H10) *holds. Then*

$$\liminf_{h\searrow 0} \frac{V_C(\Delta_0+h)-V_C(\Delta_0)}{h} \geq \inf_{x\in\tilde{\Sigma}_C(\Delta_0)} \inf_{(w,p,q)\in\tilde{M}^1(x,P_C(\Delta_0))} \tilde{E}(\Delta_0,x,w(b),q),$$

and if $\Delta_0 > 0$, *then*

$$\limsup_{h\nearrow 0} \frac{V_C(\Delta_0+h)-V_C(\Delta_0)}{h} \leq \sup_{x\in\tilde{\Sigma}_C(\Delta_0)} \sup_{(w,p,q)\in\tilde{M}^1(x,P_C(\Delta_0))} \tilde{E}(\Delta_0,x,w(b),q).$$

Proof. The basic idea is to replace $P_C(\Delta)$ by a free endpoint problem that satisfies the hypotheses of §5, and that has the same set of optimal solutions. The multipliers will also be related, and this will allow the conclusions of the theorem to be read off from the assertions of Theorems 5.1 and 6.1. We will only give the details for the right-hand limits and $0 < \Delta_0 < \overline{\Delta}$, the left-hand limits and the case $\Delta_0 = 0$ being similar.

It is convenient, then, to introduce further notation for the "penalized" problems. For $k > 0$, let $P_C^k(\Delta)$ denote the problem

$$\min \ell(x(b)) + kd_C(x(b)) + \int_a^b L(t,\Delta,x(t),x(t-\Delta),u(t))\,\mathrm{d}t$$

$$\text{over } u(\cdot)\in\mathscr{M}[a,b] \text{ satisfying } u(t)\in U(t) \quad \text{a.e. } t\in[a,b],$$

$$\text{where } x(\cdot)=x^{u,\Delta}(\cdot).$$

The rest of the notation is carried over to this context by adding a superscript k. That is, we have $\Sigma_C^k(\Delta), V_C^k(\Delta), P_{C+\xi}^k(\Delta)$, etc.

Step 1. As the first step, we use the normality hypothesis at Δ_0 to show that $P_{C+\xi}(\Delta)$ inherits feasibility and normality for all (Δ,ξ) near $(\Delta_0, 0)$. We begin with a controllability result of considerable independent interest.

PROPOSITION 7.2. *There exists* $\delta > 0$ *and* $0 \leq \Delta_1 < \Delta_0 < \Delta_2 < \overline{\Delta}$ *so that* $V_{C+\xi}(\Delta)$ *is finite for each* $|\xi| \leq \delta$ *and* $\Delta \in [\Delta_1, \Delta_2]$.

Proof. We argue by contradiction. Suppose that there exist sequences $\{\Delta_i\} \subseteq [0,\overline{\Delta})$, $\{\xi_i\} \subseteq \mathbb{R}^n$ so that $\Delta_i \to \Delta_0$, $\xi_i \to 0$ as $i \to \infty$, and $V_{C+\xi_i}(\Delta_i) = +\infty$ for all i. Define the reachable set $R(\Delta)$ by

$$R(\Delta) := \{x(b) : x(\cdot) = x^{u,\Delta}(\cdot), \quad u(\cdot) \text{ admissible}\}.$$

The assumptions $V_C(\Delta_0) < \infty$ and $V_{C+\xi_i}(\Delta_i) = +\infty$ are then equivalent to $R(\Delta_0) \cap C \neq \emptyset$ and $R(\Delta_i) \cap (C+\xi_i) = \emptyset$. Let $u_0(\cdot) \in \Sigma_C(\Delta_0)$, and so $x^{u_0,\Delta_0}(b) \in R(\Delta_0) \cap C$. As shown in §8A, the map $\Delta \to x^{u_0,\Delta}(b)$ is continuous at Δ_0, and therefore $d_{C+\xi_i}(x^{u_0,\Delta_i}(b)) \longrightarrow 0$ as $i \to \infty$. Let $\{k_i\}$ be a sequence such that $k_i \nearrow +\infty$ and $k_i d_{C+\xi_i}(x^{u_0,\Delta_i}(b)) \longrightarrow 0$.

Now let $u_i(\cdot) \in \sum_{C+\xi_i}^{k_i}(\Delta_i)$ and $x_i(\cdot) = x^{u_i,\Delta_i}(\cdot)$. We claim that $d_C(x_i(b)) \longrightarrow 0$ as $i \to \infty$. Indeed, if not, since (H5) and (H6) imply $J(\Delta,x,u)$ is bounded over all $\Delta \in [0,\overline{\Delta})$, admissible $u(\cdot)$, and $x(\cdot) = x^{u,\Delta}(\cdot)$, it would follow that $\limsup_{i\to\infty} V_{C+\xi_i}^{k_i}(\Delta_i) = +\infty$. On the other hand,

$$\begin{aligned}\limsup_{i\to\infty} V_{C+\xi_i}^{k_i}(\Delta_i) &\leq \limsup_{i\to\infty}\{J(\Delta_i,x^{u_0,\Delta_i}(\cdot),u_0) + k_i d_{C+\xi_i}(x^{u_0,\Delta_i}(b))\} \\ &= J(\Delta_0,x^{u_0,\Delta_0},u_0) = V_C(\Delta_0).\end{aligned} \tag{7.2}$$

This is a contradiction, which verifies the claim. We pass to a subsequence if necessary so that $x_i(b) \to \pi$ for some $\pi \in C$. Furthermore, we may assume that $x_i(\cdot)$ converges uniformly to some $x(\cdot)$, and $\dot{x}_i(\cdot)$ converges weakly in $L^2[a,b]$ to $\dot{x}(\cdot)$. We next claim that $x(\cdot) \in \tilde{\Sigma}_C(\Delta_0)$. To see this we can show, as in the proof of Lemma 6.2, that there exists admissible $v(\cdot)$ with $x(\cdot) = x^{v,\Delta_0}(\cdot)$, and such that

$$J(\Delta_0, x, v) \leq \liminf_{i\to\infty} J(\Delta_i, x_i, u_i).$$

Now $u_i \in \Sigma^{k_i}_{C+\xi_i}(\Delta_i)$, and so we conclude from (7.2) that

$$\limsup_{i\to\infty} J(\Delta_i, x_i, u_i) \leq \limsup_{i\to\infty} V^{k_i}_{C+\xi_i}(\Delta_i) \leq V_C(\Delta_0).$$

Hence $J(\Delta_0, x, v) \leq V_C(\Delta_0)$, and since $x(b) \in C$, it follows that $x(\cdot) \in \tilde{\Sigma}_C(\Delta_0)$, and the claim is proven. We will arrive at a contradiction by showing that $x(\cdot)$ is not normal.

Let $\pi_i \in C$ be chosen so that $d_{C+\xi_i}(x_i(b)) = |x_i(b) - \xi_i - \pi_i|$. Since $x_i(\cdot) \in \tilde{\Sigma}^{k_i}_{C+\xi_i}(\Delta_i)$, it follows readily that $x_i(\cdot)$ is also an optimal trajectory for the problem $P^{k_i}_{\{\xi_i+\pi_i\}}(\Delta_i)$. The necessary conditions guarantee the existence of arcs (w_i, p_i, q_i) contained in $\tilde{M}\left(x_i, P^{k_i}_{\{\xi_i+\pi_i\}}(\Delta_i)\right)$, which means that

$$(7.3) \qquad w_i(a) = 0, \quad -p_i(b) \in \hat{\partial}\ell(x_i(b)) + k_i\hat{\partial}d_{\{\xi_i+\pi_i\}}(x_i(b)), \quad q_i(b) = 0$$

and

$$(7.4) \quad (-\dot{w}_i(t), -\dot{p}_i(t), -\dot{q}_i(t), \dot{x}_i(t)) \in \partial H^1(t, \Delta_i, x_i(t), x_i(t-\Delta_i), p_i(t) + q_i(t+\Delta_i)) \quad \text{a.e. } t \in [a,b].$$

Since $x_i(b) - \xi_i \neq \pi_i$ for all i, $\zeta_i := (x_i(b) - \xi_i - \pi_i/|x_i(b) - \xi_i - \pi_i|)$ satisfies $|\zeta_i| = 1$ and $\{\zeta_i\} = \hat{\partial}d_{\{\xi_i+\pi_i\}}(x_i(b))$. Now elements of $\hat{\partial}\ell(x_i(b))$ are contained in a bounded set independent of i, and since $k_i \to \infty$, it follows from (7.3) that $|p_i(b)| \to +\infty$. Set $\lambda_i = |p_i(b)|^{-1}$ and define $(\overline{w}_i(t), \overline{p}_i(t), \overline{q}_i(t)) = \lambda_i(w_i(t), p_i(t), q_i(t))$. We have

$$(7.5) \qquad \overline{w}_i(a) = 0, \quad -\overline{p}_i(b) \in \lambda_i\hat{\partial}\ell(x_i(b)) + k_i\lambda_i\zeta_i, \quad \overline{q}_i(b) = 0.$$

Also by [C1, Prop. 3.2.4 (e)] and (7.4), we have

$$(7.6) \quad (-\dot{\overline{w}}_i(t), -\dot{\overline{p}}_i(t), -\dot{\overline{q}}_i(t), \dot{x}_i(t)) \in \partial H^{\lambda_i}(t, \Delta_i, x_i(t), x_i(t-\Delta_i), \overline{p}_i(t) + \overline{q}_i(t+\Delta_i)) \quad \text{a.e. } t \in [a,b].$$

Again we pass to a subsequence of $(\overline{w}_i, \overline{p}_i, \overline{q}_i)$, and obtain (w,p,q) so that $(\dot{\overline{w}}_i, \dot{\overline{p}}_i, \dot{\overline{q}}_i) \to (\dot{w}, \dot{p}, \dot{q})$ weakly in $L^1_{2n+1}[a,b]$ and $(\overline{w}_i, \overline{p}_i, \overline{q}_i) \to (w,p,q)$ uniformly on $[a,b]$. Obviously, $|p(b)| = 1$, and $-p(b) = \lim_{i\to\infty} \zeta_i$. Now $\zeta_i \in \hat{\partial}d_C(\pi_i)$ and $\pi_i \to \pi = x(b)$, hence $-p(b) \in \hat{\partial}d_C(x(b))$ by (H1$''$). Also (7.5) has $w(a) = 0$ and $q(b) = 0$. Furthermore, as we have used many times previously, (7.6) is preserved in the limit. That is,

$$(7.7) \quad (-\dot{w}(t), -\dot{p}(t), -\dot{q}(t), \dot{x}(t)) \in \partial H^0(t, \Delta_0, x(t), x(t-\Delta_0), p(t) + q(t+\Delta_0)) \quad \text{a.e. } t \in [a,b].$$

We conclude that $(w,p,q) \in \tilde{M}^0(x, P_C(\Delta_0))$ with $p \neq 0$, and this contradicts the normality of $x(\cdot)$. □

LEMMA 7.3. *There exists $\delta > 0$ and $0 \leq \Delta_1 < \Delta_0 < \Delta_2 < \overline{\Delta}$ so that $P_{C+\xi}(\Delta)$ is normal for each $|\xi| < \delta$ and $\Delta \in [\Delta_1, \Delta_2]$.*

Proof. For the proof, assume otherwise. There exist sequences $\{\Delta_i\}$, $\{\xi_i\}$ (with $\Delta_i \to \Delta_0, \xi_i \to 0$ as $i \to \infty$) such that for each i, there exist nonnormal arcs $x_i(\cdot) \in \tilde{\Sigma}_{C+\xi_i}(\Delta_i)$ and $(w_i, p_i, q_i) \in \tilde{M}^0(x_i, P_{C+\xi_i}(\Delta_i))$ with $p_i \not\equiv 0$. We may assume $|p_i(b)| = 1$. Once again, without relabeling, we pass to the convergence of subsequences, and obtain (w,p,q,x) for which $(\dot{w}_i, \dot{p}_i, \dot{q}_i, \dot{x}_i) \to (\dot{w}, \dot{p}, \dot{q}, \dot{x})$ weakly in $L^1_{3n+1}[a,b]$ and $(w_i, p_i, q_i, x_i) \to (w,p,q,x)$ uniformly on $[a,b]$. Previous arguments show that $x(\cdot) \in \tilde{\Sigma}_C(\Delta_0)$ and (w,p,q,x) satisfies the Hamiltonian inclusion (7.7). To obtain the contradiction $(w,p,q) \in \tilde{M}^0(x, P_C(\Delta_0))$, $p \neq 0$, it is only left to show that $-p(b) \in \hat{N}_C(x(b))$.

Since $-p_i(b) \in \operatorname{cl} \bigcup_{k>0} k\, \hat{\partial} d_{C+\xi_i}(x_i(b))$, there exists $k_i > 0$ and $\zeta_i \in \hat{\partial} d_{C+\xi_i}(x_i(b))$ so that $|-p_i(b) - k_i\zeta_i| \longrightarrow 0$ as $i \to \infty$. Without loss of generality, we may take $|\zeta_i| = 1$ for all i, which implies $k_i \to 1$. Note that $\hat{\partial} d_{C+\xi_i}(x_i(b)) = \hat{\partial} d_C(x_i(b) - \xi_i)$ and $x_i(b) - \xi_i \longrightarrow x(b)$. Therefore, the upper semicontinuity of $\hat{\partial} d_C$ gives that $-p(b) = \lim_{i\to\infty} \zeta_i \in \hat{\partial} d_C(x(b))$. This finishes the proof that $x(\cdot)$ is not normal, a contradiction at which the proof of Lemma 7.3 is complete. □

Henceforth, we fix $\delta > 0$ and $0 \leq \Delta_1 < \Delta_0 < \Delta_2 < \overline{\Delta}$ so that Proposition 7.2 and Lemma 7.3 are valid.

Step 2. In this step, we will obtain bounds on a Lipschitz constant that will be uniform in Δ. It is convenient to alter our notation *in this step only*, and write $V(\Delta, \xi)$ for $V_{C+\xi}(\Delta)$. We recall the following sensitivity estimate from [CW].

PROPOSITION 7.4 ([CW, Thm. 4]). *Suppose that $V(\Delta, \xi)$ is finite and $P_{C+\xi}(\Delta)$ is normal. Then $\xi' \longrightarrow V(\Delta, \xi')$ is Lipschitz near ξ with*

$$\partial_\xi V(\Delta, \xi) \subseteq \operatorname{cl}\ \operatorname{co} \Big\{ \zeta + p(b) : x(\cdot) \in \tilde{\Sigma}_{C+\xi}(\Delta), \zeta \in \hat{\partial}\ell(x(b)), \\ (w,p,q) \in \tilde{M}^1(x, P_{C+\xi}(\Delta)) \Big\}.$$

We next show that $\partial_\xi V(\Delta, \xi)$ is bounded independent of $\Delta \in [\Delta_1, \Delta_2]$ and $|\xi| \leq \delta$. Again, normality is the key ingredient.

LEMMA 7.5. *Define*

$$k_0 := \sup_{\substack{\Delta_1 \leq \Delta \leq \Delta_2 \\ |\xi| \leq \delta}} \{|z| : z \in \partial_\xi V(\Delta, \xi)\}.$$

Then $k_0 < +\infty$.

Proof. By Step 1 and Proposition 7.4, we have for each fixed $\Delta \in [\Delta_1, \Delta_2]$ and $|\xi| \leq \delta$ that $\sup\{|z| : z \in \partial_\xi V(\Delta, \xi)\} < \infty$. So if $k_0 = \infty$, then by Proposition 7.4 there exist $\Delta_i \in [\Delta_1, \Delta_2]$, $|\xi_i| \leq \delta$, $x_i(\cdot) \in \tilde{\Sigma}_{C+\xi_i}(\Delta_i)$, $\zeta_i \in \hat{\partial}\ell(x_i(b))$ and $(w_i, p_i, q_i) \in \tilde{M}^1(x_i, P_{C+\xi_i}(\Delta_i))$ such that $|\zeta_i + p_i(b)| \longrightarrow \infty$ as $i \to \infty$. It is immediate from (H5) that $\{x_i(b)\}$ is bounded, whence from (H1$''$) $\{\zeta_i\}$ is also bounded. It must therefore be the case that $\lambda_i := |p_i(b)|^{-1} \to 0$. The argument proceeds along now familiar lines: we may assume $\Delta_i \longrightarrow \Delta^* \in [\Delta_1, \Delta_2]$ and $\xi_i \to \xi^*$ with $|\xi^*| \leq \delta$. Set $(\overline{w}_i, \overline{p}_i, \overline{q}_i) = \lambda_i(w_i, p_i, q_i)$. Passing to an appropriate subsequence, we obtain (w,p,q,x) for which $(\dot{\overline{w}}_i, \dot{\overline{p}}_i, \dot{\overline{q}}_i, \dot{x}\cdot) \to (\dot{w}, \dot{p}, \dot{q}, \dot{x})$ weakly in $L^1_{3n+1}[a,b]$ and $(\overline{w}_i, \overline{p}_i, \overline{q}_i, x_i) \to (w,p,q,x)$ uniformly on $[a,b]$. We can argue as earlier (see the proof of Proposition 7.2) that $x(\cdot) \in \tilde{\Sigma}_{C+\xi^*}(\Delta^*)$ and $(w,p,q) \in \tilde{M}^0(x, P_{C+\xi^*}(\Delta^*))$ with $|p(b)| = 1$. This contradicts the normality of $P_{C+\xi^*}(\Delta^*)$ that was shown to hold in Lemma 7.3. □

Step 3. We are now ready to begin replacing $P_C(\Delta)$ by a free endpoint problem. The relevance of Steps 1 and 2 is that the penalization parameter can be chosen independently of $\Delta \in [\Delta_1, \Delta_2]$ and $|\xi| \leq \delta/2$.

LEMMA 7.6. *There exists $k_1 > 0$ so that for all $\Delta \in [\Delta_1, \Delta_2]$, $|\xi| \leq \delta/2$, and $u(\cdot) \in \Sigma^{k_1}_{C+\xi}(\Delta)$, we have $d_{C+\xi}(x^{u,\Delta}(b)) \leq \delta/2$.*

Proof. As noted earlier, (H5) and (H6) imply that there exists $\gamma > 0$ so that $|J(\Delta, x, u)| \leq \gamma$ for all $\Delta \in [0, \overline{\Delta})$, admissible $u(\cdot)$, and $x(\cdot) = x^{u,\Delta}(\cdot)$. Consequently, since $V_{C+\xi}(\Delta)$ is finite by Proposition 7.2, we have

$$\sup_{\substack{|\xi|\leq\delta/2 \\ \Delta\in[\Delta_1,\Delta_2]}} V_{C+\xi}(\Delta) \leq \gamma.$$

Let $k_1 > (4\gamma/\delta)$. Now suppose $|\xi| \leq \delta/2$, $\Delta \in [\Delta_1, \Delta_2]$, $u(\cdot) \in \Sigma^{k_1}_{C+\xi}(\Delta)$, and $x(\cdot) = x^{u,\Delta}(\cdot)$. If $d_{C+\xi}(x(b)) > \delta/2$, then

$$\begin{aligned} V^{k_1}_{C+\xi}(\Delta) &= J(\Delta, x, u) + k_1 d_{C+\xi}(x(b)) \\ &\geq -\gamma + k_1\,\delta/2 > \gamma \\ &\geq V_{C+\xi}(\Delta). \end{aligned} \tag{7.8}$$

For any $k > 0$, however, it is always the case that

$$V^{k}_{C+\xi}(\Delta) \leq V_{C+\xi}(\Delta). \tag{7.9}$$

Since (7.8) and (7.9) cannot both be valid, we conclude that $d_{C+\xi}(x(b)) \leq \delta/2$. □

LEMMA 7.7. *For $k > \max\{k_0, k_1\}$, we have $\Sigma^{k}_{C+\xi}(\Delta) = \Sigma_{C+\xi}(\Delta)$ for all $|\xi| \leq \delta/2$ and $\Delta \in [\Delta_1, \Delta_2]$.*

Proof. Fix $k > \max\{k_0, k_1\}$, $\Delta \in [\Delta_1, \Delta_2]$, and $|\xi| \leq \delta/2$. It suffices to show that if $u \in \Sigma^{k}_{C+\xi}(\Delta)$, then $x^{u,\Delta}(b) \in C + \xi$.

Let $u \in \Sigma^{k}_{C+\xi}(\Delta)$ and $x(\cdot) = x^{u,\Delta}(\cdot)$. We have

$$\begin{aligned} &V_{\{x(b)\}}(\Delta) + k\, d_{C+\xi}(x(b)) \\ &\quad\leq J(\Delta, x, u) + k\, d_{C+\xi}(x(b)) \\ &\quad= V^{k}_{C+\xi}(\Delta) \\ &\quad\leq V_{C+\xi}(\Delta) \qquad \text{(by (7.9))}. \end{aligned} \tag{7.10}$$

Let $\pi \in C + \xi$ such that $|\pi - x(b)| = d_{C+\xi}(x(b))$. By Lemma 7.6, we have $|\pi - x(b)| \leq \delta/2$. Also, a consequence of Lemma 7.5 is that $\xi' \to V_{C+\xi'}(\Delta)$ is Lipschitz of order k_0 on $|\xi'| \leq \delta$. Hence

$$\begin{aligned} V_{C+\xi}(\Delta) &\leq V_{C+\xi+x(b)-\pi}(\Delta) + k_0|x(b) - \pi| \\ &\leq V_{\{x(b)\}}(\Delta) + k_0 d_{C+\xi}(x(b)), \end{aligned} \tag{7.11}$$

the last inequality valid since $x(b) \in C + \xi + x(b) - \pi$. Combining (7.10) and (7.11) gives $k\, d_{C+\xi}(x(b)) \leq k_0 d_{C+\xi}(x(b))$, which can only be valid if $x(b) \in C + \xi$. □

Fix $k > 0$ so that Lemma 7.7 holds. The following corollary is immediate from Lemma 7.7 and the definitions.

COROLLARY 7.8. *For all $\Delta \in [\Delta_1, \Delta_2]$ and $|\xi| \leq \delta/2$, we have $V^{k}_{C+\xi}(\Delta) = V_{C+\xi}(\Delta)$ and $\tilde{\Sigma}^{k}_{C+\xi}(\Delta) = \tilde{\Sigma}_{C+\xi}(\Delta)$.*

Step 4. It remains only to compare multipliers and to apply previous results.

LEMMA 7.9. *Fix $\Delta \in [\Delta_1, \Delta_2]$ and $|\xi| < \delta/2$. Let $u(\cdot) \in \Sigma^k_{C+\xi}(\Delta)$ and $x(\cdot) = x^{u,\Delta}(\cdot)$. Then we have*

$$M(u, P^k_{C+\xi}(\Delta)) \subseteq M^1(u, P_{C+\xi}(\Delta)) \tag{7.12}$$

and

$$\tilde{M}(x, P^k_{C+\xi}(\Delta)) \subseteq \tilde{M}^1(x, P_{C+\xi}(\Delta)). \tag{7.13}$$

Proof. The only difference between the multipliers of the free endpoint problem $P^k_{C+\xi}(\Delta)$ and the constrained problem $P_{C+\xi}(\Delta)$ occurs in the transversality condition on $p(\cdot)$. The former has

$$-p(b) \in \hat{\partial}\ell(x(b)) + k\,\hat{\partial} d_{C+\xi}(x(b)).$$

However, since $\hat{\partial}\ell(x(b)) + k\,\hat{\partial} d_{C+\xi}(x(b)) \subseteq \hat{\partial}\ell(x(b)) + \hat{N}_{C+\xi}(x(b))$, it follows that the transversality condition defining $M^1(x, P_{C+\xi}(\Delta))$ is also satisfied. Hence (7.12) holds. The same reasoning shows that (7.13) is valid as well. □

End of proof. To verify the assertions of the theorem, we now need only to apply previous results to the problem $P^k_C(\Delta)$. Since the value functions and optimal solutions of $P^k_C(\Delta)$ and $P_C(\Delta)$ coincide for all Δ near Δ_0 (Lemma 7.7, Corollary 7.8), and the multiplier set of $P_C(\Delta_0)$ is larger than that of $P^k_C(\Delta_0)$ (Lemma 7.9), the bounds in Theorem 7.1(a) follow immediately from Theorem 5.1(a). Assumption (H8) on $P_C(\Delta_0)$ carries over to the same assumption on $P^k_C(\Delta_0)$ by Lemma 7.7, hence (b) follows from Theorem 5.1(b). Finally, (H9) and (H10) are only assumptions on the Hamiltonian and are not affected by the endpoint constraints. Thus (c) follows from Theorem 6.1. □

Remarks. (i) The proof of Theorem 7.1 shows that the multipliers for $P_C(\Delta)$ can be replaced by multiplier sets of the form (for k sufficiently large) :

$$M^\lambda_k(u, P_C(\Delta)) = \left\{ (p,q) : -p(b) \in \lambda\,\hat{\partial}\ell(x(b)) + k\,\hat{\partial} d_C(x(b)),\right.$$
$$q(b) = 0, \quad (p,q) \text{ satisfy the adjoint}$$
$$\left.\text{equations and the maximum condition}\right\}.$$

Similarly for $\tilde{M}^\lambda_k(x, P_C(\Delta))$. This gives potentially finer estimates.

(ii) The statement of Theorem 7.1 does not allow for ℓ to depend explicitly on Δ. The reason is that to apply Theorem 5.1, we must have the subgradient of the endpoint cost function with respect to the state variable to be upper semicontinuous. That is, $(\Delta,\xi) \rightrightarrows \hat{\partial}_\xi(\ell(\Delta,\xi) + d_C(\xi))$ must have a closed graph. This is not necessarily true if ℓ only satisfies (H1′). It is true, however, if ℓ satisfies (H1) (see [C1, p. 39]). Hence Theorem 7.1 could be given alternatively under (H1) rather than (H1″). In fact, (H1) could be weakened further to (H1′) plus "regularity" (see [C1]).

The relevant analogue of (H11) to endpoint constrained problems is that $M^1(x, P_C(\Delta_0))$ consists of a single element. We will encounter in §8C an example where this is so. The following corollary is the endpoint constrained version of Corollary 6.5.

COROLLARY 7.10. *Fix $\Delta_0 \in [0, \overline{\Delta})$. In addition to the basic hypotheses of Theorem 7.1, suppose that (H11) and either (H9) or (H10) hold. Then*

(a) $\lim_{h\searrow 0} \dfrac{V(\Delta_0+h)-V(\Delta_0)}{h} = \inf_{v\in\Sigma_C(\Delta_0)} E(\Delta_0, x^{v,\Delta_0}, v, (p,q)),$
where $M^1(v, P_C(\Delta_0)) = \{(p,q)\}$*;*

(b) *If* $\Delta_0 > 0$*, then*

$$\lim_{h\nearrow 0} \frac{V(\Delta_0+h)-V(\Delta_0)}{h} = \sup_{v\in\Sigma_C(\Delta_0)} E(\Delta_0, x^{v,\Delta_0}, v, (p,q)),$$

where $M^1(v, P_C(\Delta_0)) = \{(p,q)\}$*;*

(c) *Suppose, in addition, that* $\Sigma_C(\Delta_0)$ *consists of a single element* v*. Let* $M^1(v, P(\Delta_0)) = \{(p,q)\}$*. Then if* $\Delta_0 > 0$*, the derivative of* V *at* Δ_0 *exists and equals* $E(\Delta_0, x^{v,\Delta_0}, v, (p,q))$*.*

Proof. Note that (H11) implies that for $v \in \Sigma_C(\Delta_0), M^1(v, P_C(\Delta_0))$ also consists of a single element, and is the (p,q) coordinate of $\tilde{M}^1(x^{v,\Delta_0}, P_C(\Delta_0))$. (See the remark at the end of §6.) Moreover, in this case, the E and $\tilde{E}$ functionals give the same value, hence the corollary is an immediate consequence of Theorem 7.1(a) and (c). □

8. Three examples.

A. Endpoint dependence of a differential-difference equation. Consider the (uncontrolled) differential-difference equation

$$\text{(8.1)}\qquad \begin{aligned} \dot{x}(t) &= \phi(t, x(t), x(t-\Delta)) \text{ a.e. } t \in [a,b], \\ x(t) &= c(t) \text{ for } t \in [a-\Delta, a]. \end{aligned}$$

Assume that ϕ satisfies (H1)–(H6), except that now it is independent of $u \in \mathbb{R}^m$ and $\Delta \in [0, \bar{\Delta})$. The existence theory for (8.1) proclaims that a unique solution $x^\Delta(\cdot)$ exists, but as mentioned in the Introduction, little seems to be known concerning the dependence of $x^\Delta(\cdot)$ on the delay Δ. We show as an application of the results of §3 that $\Delta \to x^\Delta(b)$ is differentiable on $(0, \bar{\Delta})$, and present the precise formula for its right derivative at $\Delta = 0$. This result appears to be new.

PROPOSITION 8.1. *Define* $f : [0, \overline{\Delta}) \to \mathbb{R}^n$ *by* $f(\Delta) = x^\Delta(b)$*. Then for each* $\Delta \in (0, \overline{\Delta})$*,* $d/d\Delta f(\Delta)$ *exists. Moreover, the right derivative at* 0 *is given by*

$$\text{(8.2)}\qquad \frac{d^+}{d\Delta} f(0) = \int_a^b \phi_2(t, x^0(t), x^0(t))^*(Q(t))\dot{x}^0(t)\,dt,$$

where $Q(\cdot)$ *is the solution of the* $n \times n$ *matrix differential equation*

$$-\dot{Q}(t) = (\phi_1^*(t, x^0(t), x^0(t)) + \phi_2^*(t, x^0(t), x^0(t)))Q(t) \quad \text{a.e. } t \in [a,b],$$

$$-Q(b) = I \quad (= \text{identity matrix}).$$

Proof. Fix $j \in \{1, 2, \cdots, n\}$. Write $f^{(j)}(\Delta)$ for the jth component of $f(\Delta)$. We show $f^{(j)}(\Delta)$ is differentiable at each $\Delta \in (0, \Delta_0)$.

In the problem formulation $P(\Delta)$, define $\ell : [0, \overline{\Delta}) \times \mathbb{R}^n \longrightarrow \mathbb{R}^1$ by $\ell(\Delta, \xi) = \xi^{(j)}$, where $\xi^{(j)}$ denotes the jth component of ξ, and set $L \equiv 0$. Then $V(\Delta) = f^{(j)}(\Delta)$, and it is immediate from Corollary 3.3 that $V(\Delta)$ is differentiable for $\Delta \in (0, \overline{\Delta})$. (Equation (H8) is satisfied trivially.) Hence f is differentiable at each $\Delta \in (0, \overline{\Delta})$.

Of course, the above argument is also applicable when $\Delta = 0$. Now if (p,q) is the multiplier for $P(0)$ with this data, then $-p(b) = e_j$ (= the column vector with 1 in the jth coordinate and zeros elsewhere) and $q(b) = 0$, and $-(\dot{p}(t) + \dot{q}(t)) =$

$(\phi_1(t,x(t),x(t))+ \phi_2^*(t,x(t),x(t)))\,(p(t)+q(t))$ almost everywhere $t \in [a,b]$. Therefore $p(t)+q(t)$ is the jth column of $Q(t)$. Now by Theorem 3.2(a),

$$\frac{d^+}{d\Delta}V(0) = \int_a^b \langle -\dot q(t), \dot x^0(t)\rangle \,\mathrm{dt}.$$

Also, we have $-\dot q(t) = \phi_2^*(t,x^0(t),x^0(t))(p(t)+q(t))$. Hence $\lim_{\Delta\downarrow 0}(f^{(j)}(\Delta) - f^{(j)}(0)/\Delta)$ exists and equals the jth column of (8.2). This is true for each j, so the proof is complete. □

B. The linear problem of Mayer. We now study the case in which $\phi(t,\Delta,x,y,u)$ has the form $Ax+By+Cu$, where A and B are given $n\times n$ matrices and C a given $n\times m$ matrix. We suppose, in addition, that L is identically zero, U is a compact convex subset of $\mathbb{R}^m$, and ℓ satisfies (H1). Consider the (free endpoint) problem $P(\Delta_0)$ of §6. It is easy to confirm that (H1)–(H7) are satisfied, as well as (H11) (in the present setting, there is a one-to-one correspondence between Hamiltonian multipliers and Pontryagin multipliers).

The Hamiltonian H is given by

$$\begin{aligned} H(t,\Delta,x,y,p) &= \max_{u\in U}\langle p, Ax+By+Cu\rangle \\ &= \langle p, Ax+By\rangle + \max_{u\in U}\langle p, Cu\rangle. \end{aligned}$$

It is clear from this expression that (H9) is satisfied. With the preceding, we can derive from Corollary 6.5 a rather explicit formula for $(d^+/d\Delta)V(0)$.

PROPOSITION 8.2. *V is differentiable on $(0,\overline{\Delta})$. Moreover, the right derivative $(d^+/d\Delta)V(0)$ is given by*

$$\frac{d^+}{d\Delta}V(0) = \inf_{v\in\Sigma(0)} \frac{\partial}{\partial\Delta}\ell(\Delta,x(b)) + \int_a^b \langle -\dot q(t), \dot x(t)\rangle\,\mathrm{dt},$$

where

$$x(t) = e^{(A+B)(t-a)}c(a) + \int_a^t e^{(A+B)(t-s)}Cv(s)\,\mathrm{ds}$$

and

$$q(t) = B^* e^{(A^*+B^*)(b-t)}(-\nabla_x \ell(0,x(b))).$$

If $AB=BA$, the formula for $(d^+/d\Delta)V(0)$ further simplifies to

$$\inf_{v\in\Sigma(0)} \frac{\partial}{\partial\Delta}\ell(0,x(b)) + \langle \zeta, D(a)c(a) + \int_a^b D(s)Cv(s)\,\mathrm{ds}\rangle,$$

where $x(\cdot)$ is as above, $\zeta = -\nabla_x\ell(0,x(b))$, and D is the map from $[a,b]$ into the $n\times n$ matrices given by

$$D(s) = (b-s)e^{(A+B)(b-s)}B.$$

C. An example in renewable resource modeling. Our final example treats a more specific problem involving one of the better-known models in resource theory [C]. It involves a population biomass $x(t)$ evolving according to the dynamic equation

$$\dot x(t) = g(x(t)) - u(t)x(t), \qquad x(0) = x_0, \tag{8.3}$$

where g, the natural growth function, is C^1 and concave, x_0 is given, and the last term on the right in the dynamics reflects the effect on the population of applying a harvesting effort u (assumed to be constrained to a given interval $[0, u_{\max}]$). The discounted net revenue resulting from a choice of effort profile $u(t)$ is given by

$$\int_0^T e^{-\delta t}\{\pi x(t) - c\}u(t)dt, \tag{8.4}$$

where T is the planning horizon, δ the discount rate, π the resource price, and c the effort cost. We will impose the constraint $x(T) = x_T$ (given). The issue is to identify the control $u(\cdot)$ that maximizes revenue, subject to the constraints.

The nondelay problem described above is well understood. Under standard assumptions that we omit, the unique solution is of "turnpike" type; we briefly describe it now. A certain target population level x^* is identified by an algebraic equation involving all the data except the endpoint values. The optimal solution employs either no harvesting effort ($u = 0$) or maximal effort ($u = u_{\max}$) to guide $x(t)$ to x^* at the start (depending on whether x_0 is less than or greater than x^*). In the intermediate (singular or turnpike) stage, precisely the effort required to stay at x^* is applied. At the end, to honour the endpoint constraint, there is an interval in which u is constant: either 0 (if $x_T > x^*$) or $u_{\max}$ (if $x_T < x^*$). The two switching times will be denoted τ_1 and τ_2.

Determining x^*, τ_1, and τ_2 precisely is done by means of the (nondelay) maximum principle. The adjoint variable $r(t)$ is unique and positive throughout, and satisfies the adjoint equation

$$-\dot{r}(t) = [g'(x(t)) - u(t)]r(t) + e^{-\delta t}\pi u(t),$$

and the condition

$$u(t) \text{ maximizes } u \rightarrow \{e^{-\delta t}(\pi x(t) - c) - r(t)x(t)\}u \text{ over } u \in [0, u_{\max}].$$

Suppose now that we wish to consider the possible effect of a small delay in the natural growth law, a biologically realistic possibility. Then (8.3) might become, for example,

$$\dot{x}(t) = g(x(t - \Delta)) - u(t)x(t), \tag{8.5}$$

with $x(t)$ specified on $[-\Delta, 0]$.

The resulting time delay optimal control problem is no longer amenable to analytical solution; in fact, the solution is not known. We will apply the results of §7 to calculate the marginal effect of a small delay. (Hypotheses (H1)–(H7) are all satisfied under the usual assumptions.)

The Hamiltonian $\mathscr{H}$ for the new problem (8.5) is given by

$$\mathscr{H}(t, \Delta, x, y, u, p) = pg(y) - pux + e^{-\delta t}\{\pi x - c\}u,$$

so that (H9) holds. Now any element (w, p, q) of $\tilde{M}^1(x, P_C(0))$ is such that $w \equiv 0$ and (p, q) satisfies

$$-\dot{p}(t) = -(p(t) + q(t))u(t) + e^{-\delta t}\pi u(t),$$
$$-\dot{q}(t) = (p(t) + q(t))g'(x(t)), \quad q(T) = 0,$$

together with the maximum condition

$$u(t) \text{ maximizes } u \rightarrow \mathscr{H}(t, \Delta, x(t), x(t), u, p(t) + q(t)) \text{ over } u \in \ [0, u_{\max}].$$

We deduce from this that $p(t) + q(t)$ satisfies the (nondelay) maximum principle, whence $p + q$ and r coincide. Thus $p(T) = r(T)$ (known), and it follows that there is a single element $(0, p, q)$ in $\tilde{M}^1(x, P_C(0))$; i.e., (H11) is satisfied at $\Delta = 0$.

We are now in a position to apply Corollary 7.10, which gives a formula for $(d^+/d\Delta)V(0)$. Note that $\dot{x}$ is 0 along the intermediate stage of the solution, while $-\dot{q}(t) = r(t)g'(x(t))$ throughout. We may therefore summarize as follows.

PROPOSITION 8.3. *The following condition holds*:

$$\frac{d^+}{d\Delta}V(0) = \int_0^{\tau_1} r(t)g'(x(t))\dot{x}(t)dt + \int_{\tau_2}^{T} r(t)g'(x(t))\dot{x}(t)dt.$$

The two integrals figuring in this formula may be positive or negative in general. For example, in the typical situation in which we have $g'(x^*) > 0$, with $x_0 < x^*, x_T < x^*$, it is easy to see that the first integral is positive and the second negative.

In consequence, by setting either x_0 or x_T sufficiently near x^*, we can arrange to make $(d^+/d\Delta)V(0)$ either positive or negative, which shows that its sign cannot generally be determined by qualitative arguments.

REFERENCES

[B] H. T. BANKS, *Necessary condition for problems with variable timelags*, SIAM J. Control Optim., 6 (1968), pp. 9–47.

[BC1] R. BELLMAN AND K. COOKE, *On the limit of solutions of differential-difference equations as the retardation approaches zero*, Proc. Nat. Acad. Sci., 45 (1959), pp. 1026–1028.

[BC2] ———, *Differential-Difference Equations*, Academic Press, New York, London, 1963.

[BS] J. B. BORWEIN AND H. M. STROJWAS, *Proximal analysis and boundaries of closed sets in Banach space, Part* I: *Theory*, Canad. J. Math., 38 (1986), pp. 431-452.

[C] C. W. CLARK, *Mathematical Bioeconomics*, John Wiley, New York, 1976.

[C1] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, John Wiley, New York, 1983.

[C2] ———, *Perturbed optimal control problems*, IEEE Trans. Automat. Control, 31 (1986), pp. 535–542.

[C3] ———, *Methods of Dynamic and Nonsmooth Optimization*, CBMS–NSF Regional Conf. Ser. Appl. Math., vol. 57, Society for Industrial and Applied Mathematics., Philadelphia, PA, 1989.

[CW] F. H. CLARKE AND G. G. WATKINS, *Necessary conditions, controllability and the value function for differential-difference inclusions*, Nonlinear Anal. Theory Meth. Appl., 10 (1986), pp. 1155–1179.

[D1] R. D. DRIVER, *Some harmless delays, in Delay and Functional Differential Equations and Their Applications*, K. Schmitt, ed., Academic Press, London, 1972.

[D2] ———, *Ordinary and Delay Differential Equations*, Appl. Math. Sci., vol. 20, Springer-Verlag, New York, 1977.

[H1] A. HALANAY, *Differential Equations: Stability, Oscillations, Time Lags*, Academic Press, New York, 1966.

[H2] J. HALE, *Theory of Functional Differential Equations*, Springer-Verlag, New York, 1977.

[K] G. L. KHARATISHVILI, *Maximum principle in the theory of optimum time-delay processes*, Dokl. Akad. Nauk. USSR, 136 (1961), pp. 39–42.

[L] P. D. LOEWEN, *The proximal normal formula in Hilbert space*, Nonlinear Anal. Theory Meth. Appl., 11 (1987), pp. 979–995.

[M1] A. MANITIUS, *Optimal control of hereditary systems*, in *Control Theory and Topics in Functional Analysis*, Internat. Center Theoret. Phys., Miramare, Trieste, Italy, 1974.

[M2] ———, *On the optimal control of systems with delays depending on state, control, and time*, Research Report CRM-449, Centre de Recherches Mathématiques, Université de Montréal, 1974.

[O] M. N. OĞUZTØRELI, *Time-Lag Control Systems*, Academic Press, New York, London, 1966.

[P] L. S. PONTRYAGIN, V. G. BOLTYANSKII, R. V. GAMKRELIDZE, AND E. F. MISCHENKO, *The Mathematical Theory of Optimal Processes*, John Wiley, New York, 1962.

[S] S. SUGIYAMA, *Continuity properties on the retardation in the theory of differential-difference equations*, Proc. Japan Acad., 37 (1961), pp. 179-182.

[W1] J. WARGA, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, London, 1972.

[W2] ———, *Controllability of nondifferentiable hereditary processes*, SIAM J. Control Optim., 16 (1978), pp. 813–831.

[W3] ———, *Controllability, extremality, and abnormality in nonsmooth optimal control*, J. Optim. Theory Appl., 41(1983), pp. 239–260.

SIAM J. CONTROL AND OPTIMIZATION.
Vol. 29, No. 5, pp. 1216–1249, September 1991

A MONTE CARLO METHOD FOR SENSITIVITY ANALYSIS AND PARAMETRIC OPTIMIZATION OF NONLINEAR STOCHASTIC SYSTEMS*

JICHUAN YANG† AND HAROLD J. KUSHNER‡

Abstract. For high-dimensional or nonlinear problems there are serious limitations on the power of available computational methods for the optimization or parametric optimization of stochastic systems. The paper develops an effective Monte Carlo method for obtaining good estimators of systems sensitivities with respect to system parameters under quite general conditions on the systems and cost functions. The value of the method is borne out by numerical experiments, and the computational requirements are favorable with respect to competing methods when the dimension is high or the nonlinearities "severe." The method is a type of "derivative of likelihood ratio" method. Jump-diffusion, functional diffusion, and reflected diffusion models of broad types are covered by the basic technique (e.g., the type of limit model that arises in the analysis of queueing systems under heavy traffic, where the boundary reflection conditions are discontinuous). For a wide class of problems, the cost function or dynamics need not be smooth in the state variables; for example, where the cost is the probability of an event or "sign" functions appear in the dynamics. Under appropriate conditions, it is shown that the cost functions are differentiable with respect to the parameters. Since the basic diffusion (or other) model cannot be simulated exactly, two types of readily simulatable approximations are discussed in detail, and estimators of the derivatives of the cost functions for these approximations are obtained and analyzed. It is shown that these estimators and their expectations converge to those for the original problem. Thus, a robustness result for the sensitivity estimators, namely that the derivatives of the cost functions (and their estimators) for the simulatable approximations converge to those for the approximated process is proven. Such results are essential, in any case, if a simulation-based method is to be used with confidence.

Key words. Monte Carlo method for diffusions, parametric optimization of stochastic systems, sensitivity analysis, optimization of stochastic systems, nonlinear stochastic systems, high-dimensional stochastic systems, parametric optimization of jump-diffusion processes, likelihood ratio method for sensitivity analysis

AMS(MOS) subject classifications. 62E25, 93E20, 93E25

1. Introduction.

The field of numerical optimization (or even the design of good controls) for stochastic systems is still in its infancy. For problems where the dimension of the state is greater than 3, the methods available for computing feedback controls are too time consuming to be of use, except for some very special cases. Also, for many applications, we do not want a feedback control (a control that depends on the current value of the full state), but rather a control that depends on the observed data (say, the position, or "error" over some time interval) and is easily constructable. For such reasons, it has been common to parametrize the control in some way, then to use some "rough" analysis to get an initial value for the parameter, and finally to use some type of Monte Carlo or stochastic approximation type method and either simulation or actual operating data to improve or optimize the parameters. In this paper, we will present a novel, very interesting, and successful method for doing this on a wide variety of problems. For a large class of problems, the method has significant advantages over currently available competing methods.

* Received by the editors February 26, 1990; accepted for publication (in revised form) August 11, 1990.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The work of this author was partially supported by National Science Foundation grant NSF ECS-8505674.

‡ Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The work of this author was supported by Air Force Office of Scientific Research grant AFOSR 89-0015 and by Army Research Office grant ARO DAAL-03-86K-0171.

The basic results of the paper will be developed for systems that are fundamentally of the Itô equation type:

$$(1.1)\qquad dx = b(x,\alpha)dt + \sigma(x)dw, \quad x \in R^r, \quad \text{Euclidean r} - \text{space},$$

where α is the control parameter of concern. Jump-diffusion and various reflected diffusion models will also be dealt with. In practice, we must often work with a stochastic functional differential equation where the b and σ depend on the "past" values as well as the current value of the state. To see why this is so, consider the controlled problem modeled by the equation $dx = \tilde{b}(x,u)dt + \sigma(x)dw$, where u is the control. Commonly, we use controls that are functionals of the past observed data, say, the position $x^1(\cdot)$. Suppose that we choose the control form

$$(1.2)\qquad u(t) = \int_0^t \sum_{i=1}^{q} \beta_i e^{-\gamma_i(t-s)} x^1(s)\, ds = \sum_{i=1}^{q} u_i(t),$$

where the u_i are defined in the obvious way. Here $\alpha = \{\beta_i, \gamma_i, i \le q\}$, and the evolution equation for $x(\cdot)$ is actually a *stochastic functional differential equation.* Because of this need, most of the work in the paper uses the "functional" model. The computation is not usually as hard as it first appears. For example, for the case (1.2), $u(\cdot)$ is the solution to a linear differential equation driven by $x^1(\cdot)$. To better motivate the problem and to illustrate some points in the following sections, we describe some simple applications.

A typical application of concern is the case where the purpose of the control system is for the "position" $x^1(\cdot)$ to track a stochastic signal $z(\cdot)$, which does not depend on the control. Let E_α denote the expectation under parameter value α. For such an application, typical cost criteria of interest are

$$(1.3)\qquad V(\alpha) = E_\alpha C(x(\cdot),\alpha),$$

where, for example,

$$(1.4\text{a})\qquad C(x(\cdot),\alpha) = \int_0^T |x^1(s) - z(s)|\, ds,$$

or, alternatively,

$$(1.4\text{b})\qquad C(x(\cdot),\alpha) = I_{\{\sup_{s\le T} |x^1(s)-z(s)| \ge \Delta > 0\}}.$$

For such problems, the control is often chosen to be a linear functional of the "past" of the error $x^1(\cdot) - z(\cdot)$; e.g., for some parametrized kernel $g(\cdot)$, the control takes the form

$$(1.5)\qquad u(t) = \int_0^t g(t-s,\alpha)(x^1(s) - z(s))\, ds.$$

Another interesting example that is easily included in our framework comes from [1] and [2], which are concerned with the active control of a shock absorber of an automobile. The two-dimensional (one shock absorber) case in [1] and [2] is modeled by

$$(1.6)\qquad dx = \begin{pmatrix} x^2 \\ -ux^2 - \beta x^1 - \gamma \operatorname{sign}\ x^2 \end{pmatrix} dt + \begin{pmatrix} 0 \\ \sigma dw \end{pmatrix},$$

where $0 < \underline{u} < u \leq \overline{u} < \infty$ and

$$C(x(\cdot),\alpha) = \int_0^T |u(s)x^2(s) + \beta x^1(s) + \gamma \operatorname{sign} x^2(s)|^2 \, ds, \tag{1.7}$$

for positive constants β, γ. As a consequence of the analysis in [1], it was suggested there that suboptimal parametrized feedback controls of the "truncated" form

$$u(x,\alpha) = [\alpha_1 + \alpha_2 x^1 \text{ sign } x^2]\Big|_{\underline{u}}^{\overline{u}} \tag{1.8}$$

would work well.

The general Monte Carlo method for systems optimization involves getting a sequence of (hopefully improving) values $\{\alpha(n)\}$ of the parameter. Given the current parameter value $\alpha(n)$, we try to get a "reasonably" unbiased estimate of the gradient $V_\alpha(\alpha(n))$ and change $\alpha(n)$ according to some "stochastic gradient descent" method. One of the key questions in using Monte Carlo methods concerns how the gradient is to be estimated. Several methods are in current use, and we discuss them in §2. This paper is devoted to the development of a particularly useful and efficient "indirect" method that is based on a "derivative of a likelihood ratio," somewhat similar in spirit to [3]. The required computation scales linearly with the dimension of the state vector, plus a small additional computation that is linear in the dimension of the control. Numerical work suggests that the method is superior to the alternatives for many complex, nonlinear, or high-dimensional systems, in terms of both the variances of the estimates and the required computation time.

The general method will be defined in §3, where an unbiased estimator of $V_\alpha(\alpha)$ is given, assuming that the trajectory $x(\cdot)$ is available. In simulation work, we only have samples of *approximations of paths* of $x(\cdot)$, and not $x(\cdot)$ itself. Thus, it is important to analyze the quality of the estimates obtained by various approximations. A simple method, based on a discrete-time approximation, is given in §4. Sections 5–7 deal with a numerical method that uses a finite state Markov chain approximation. Each method has its own advantages.

In §8, we extend the result to a class of systems where the variance term also depends on the control. The general idea is applicable to any system whose "control terms" can be defined by a "Girsanov transformation." This is further illustrated in §9 by the reflected diffusion model and by a special form of the reflected diffusion which arises as a limit in the heavy-traffic modeling of a queueing system. The simulation of the physical model is usually quite hard and it is often advantageous to use the "simulatable" simple approximations to the simpler "heavy traffic" limit. In §10, we discuss the case where the noise vanishes. The basic calculations required to get the derivatives $V_\alpha(\alpha)$ for the purely deterministic problem are harder than those that our method requires for the stochastic problem. Because of this, it might be convenient to add small "artificial" noise and use the stochastic estimator. It is shown that, when a certain computable "zero mean" term is subtracted, the "vanishing noise" limit exists and is an unbiased estimator of the gradient for the deterministic problem.

The method of estimating the gradient $V_\alpha(\alpha)$ is a key component of a stochastic approximation method for optimizing the system. Such a stochastic approximation method and a detailed numerical study and comparison with alternatives will be given in a subsequent paper. Some numerical data appears in §11.

2. Some current methods for estimating $V_\alpha(\alpha)$. The discussion in this section will be somewhat loose, since we are only concerned with the relative advantages and limitations of different approaches, and it is supposed that the mathematical operations are justified. Unless otherwise mentioned, let α be real valued. The vector valued α case involves repeating the expressions for each component of α and will be commented on as necessary. Unless otherwise mentioned, we use the notation $x(\cdot,\alpha)$ for the solution of (1.1), under parameter α.

2.1. The mean square derivative method. Suppose that $C(\cdot)$ takes the form

$$C(x(\cdot,\alpha),\alpha) = \int_0^T k(x(s,\alpha),\alpha)\,ds + g(x(T,\alpha)), \tag{2.1}$$

and let $x_\alpha(\cdot,\alpha_0)$ denote the mean square derivative (assuming that it exists) of $x(\cdot,\alpha)$ with respect to α at α_0. This requires that $x(\cdot,\alpha)$ be defined with respect to the same Wiener process for α in a neighborhood of α_0. The process $x_\alpha(\cdot,\alpha_0)$ satisfies

$$\lim_{\delta\alpha\to 0} E\left|\frac{x(t,\alpha_0+\delta\alpha)-x(t,\alpha_0)}{\delta\alpha} - x_\alpha(t,\alpha_0)\right|^2 = 0,$$

for each t. For purposes of discussion, first let the model be the ordinary (not functional) Itô equation

$$dx(t,\alpha) = b(x(t,\alpha),\alpha)dt + \sigma(x(t,\alpha))dw. \tag{2.2}$$

Under differentiability conditions on the b and σ, $x_\alpha(\cdot,\alpha_0)$ satisfies [17, Chap. 8]

$$dx_\alpha(t,\alpha_0) = [b_\alpha(x(t,\alpha_0),\alpha_0) + b_x(x(t,\alpha_0),\alpha_0)\cdot x_\alpha(t,\alpha_0)]dt + dW(t), \tag{2.3}$$

where $W = (W_1,\cdots,W_r)$, and

$$dW_j = \sum_k \sigma'_{jk,x}(x(t,\alpha_0))x_\alpha(t,\alpha_0)dw_k(t),$$

and $b_x = (b_{1,x},\cdots,b_{r,x})$ is the Jacobian matrix. Then the quantity

$$\begin{aligned} Q(\alpha_0) = &\int_0^T [k_\alpha(x(s,\alpha_0),\alpha_0) + k'_x(x(s,\alpha_0),\alpha_0)x_\alpha(s,\alpha_0)]\,ds \\ &+ g'_x(x(T,\alpha_0))x_\alpha(T,\alpha_0) \end{aligned} \tag{2.4}$$

is an unbiased estimator of $V_\alpha(\alpha_0)$.

Where applicable, (2.4) is often used. For high-dimensional problems, the computational burden renders the method undesirable, since (2.2) and (2.3) must be solved. Typically, α is a vector. If α has K components, then K systems of the type (2.3) must be solved to estimate $V_\alpha(\alpha_0)$. Additionally, both the cost functional and $b(\cdot),\sigma(\cdot)$ must be x-differentiable, which eliminates cost functionals such as (1.4a), (1.4b), or systems such as (1.6). The requirement that system (2.2) must be defined with respect to the same Wiener process for α in a neighborhood of α_0 eliminates systems that are defined via the Girsanov transformation method. Also, it is difficult to extend the method to the reflected diffusion model.

If the control is of the form (1.2) or if system (2.2) is replaced by a stochastic functional differential equation, then (2.3) is replaced by a much more complicated equation, and the computational burden becomes even greater. These cited problems are not nearly as serious for the method defined in §3.

2.2. A finite difference method. Here, the estimator

$$[C(x(\cdot,\alpha_0+\delta\alpha),\alpha_0+\delta\alpha)-C(x(\cdot,\alpha_0),\alpha_0)]/\delta\alpha=\tilde{Q}(\alpha_0,\delta\alpha) \tag{2.5}$$

is used, for small $\delta\alpha$. If α is a vector with K components, then either $K+1$ (the "forward difference" method (2.5) used) or $2K$ (a "central difference" method used) simulations must be taken to get an estimator of the gradient $V_\alpha(\alpha_0)$. Thus the complexity of the computation is of the order of $r\cdot K$.

Unless special precautions are taken, the variance of the estimator goes to infinity as $\delta\alpha\to 0$. In certain cases, the variance can be kept bounded by reusing the same driving noises for all $K+1$ or $2K$ simulations, although the computational burden (for the vector-valued α case) can remain large. Suppose that for α in a neighborhood of α_0, the systems are all defined with respect to the same Wiener process. Sometimes, $x(\cdot,\alpha)$ and $C(x(\cdot,\alpha),\alpha)$ are almost everywhere continuous functions of the driving Wiener process path, and the *sample values* $C(x(\cdot,\alpha),\alpha)$ are actually differentiable functions of α at α_0. Then, we would use the same sample path of the Wiener process or an approximation to it in all the $(K+1)$ or $2K$ simulations. In practice, this approach can work quite well for small K. If the cost is of the "nonsmooth" form (1.4a), (1.4b) or if the system is highly nonlinear, then the properties deteriorate, and can be much worse than for the method developed in §3, from the points of view of the quality of the estimate and the computational requirements. This is borne out by numerical experiments.

The infinitesimal perturbation analysis (IPA) methods (of the type [4], [5]) do not seem to be applicable. In the next section, we develop the basic "likelihood ratio method," which has a reasonable computational requirement, yields estimators of "good quality" under broad conditions and avoids many of the difficulties cited above.

3. A likelihood ratio method. We will work with a class of systems for which the control terms can be defined in terms of a Girsanov measure transformation [6, Chap. IV.4]. Partition the state variable x as $x=(x_1,x_2)$, where $x_i\in R^{r_i}$, Euclidean r_i-space, and $r_1+r_2=r$. Let α_0 be interior to A_0, a compact parameter set, and let $x(\cdot)$ satisfy (3.1) under the conditions below:

$$dx=\begin{pmatrix} dx^1 \\ dx^2 \end{pmatrix}=\begin{pmatrix} b_1(x(\cdot),t) \\ b_2(x(\cdot),t,\alpha_0) \end{pmatrix}dt+\begin{pmatrix} \sigma_1(x(\cdot),t)dw_1 \\ \sigma_2(x(\cdot),t)dw_2 \end{pmatrix}. \tag{3.1}$$

α_0 will be the parameter value at which the derivative is taken. It is most convenient to do the basic work for real-valued α and then to state the trivial extensions to the vector case. Thus, unless mentioned otherwise, we suppose that α is real valued.

A remark on (3.1). Equation (3.1) looks complicated owing to the "functional" dependence of b_i and σ_i. Generally, unless the original system has a functional dependence due to, say, a delay, the dependence arises because the current value of the control might depend on past values of the state. Suppose, for example, that the model is

$$dx=b(x,u)dt+\sigma(x)dw, \tag{3.1$'$}$$

where the control u depends on α and on the "past"; e.g., for (1.2), we have

$$\dot{u}_i=\beta_i x^1-\gamma_i u_i, \tag{3.1$''$}$$

and the values of $b(\cdot)$ can be readily calculated or approximated by using (3.1″).

Let $T < \infty$ and let $C^r[0,T]$ denote the space of R^r-valued continuous functions on [0,T], with the sup norm topology. We will need the following assumptions.

(A3.1) For each $\alpha \in A_0$, $b_i(\cdot,\cdot,\alpha)$ and $\sigma_i(\cdot,\cdot)$ are measurable (vector- and matrix-valued) functions on $C^r[0,T] \times [0,T]$ and are nonanticipative in the sense that the values at time t depend only on $\{x(s), s \leq t\}$. Also, $\sigma_2^{-1}(x(\cdot),t)$ exists and is uniformly bounded. (We sometimes write $b = (b_1, b_2), \sigma = (\sigma_1, \sigma_2)$.)

(A3.2) The $w_i(\cdot), i = 1,2$, are standard vector-valued and mutually independent Wiener processes.

(A3.3) Equation (3.1) has a unique weak sense solution on $[0,T]$.

(A3.4) There is a bounded measurable function $b_{2,\alpha}(\cdot,\cdot,\alpha_0)$, which is the derivative of b_2 with respect to α at α_0 in the sense that

$$\left|\frac{b_2(x(\cdot),t,\alpha_0+\delta\alpha) - b_2(x(\cdot),t,\alpha_0)}{\delta\alpha} - b_{2,\alpha}(x(\cdot),t,\alpha_0)\right|$$

is bounded and converges to zero for almost all $(\omega,t), t \leq T$, as $\delta\alpha \to 0$.

(A3.5) For each $\alpha \in A_0, C(\cdot,\alpha)$ is a measurable function on $C^r[0,T]$. There is a measurable function $C_\alpha(\cdot,\alpha)$ such that

$$\lim_{\delta\alpha\to 0} E_{\alpha_0}\left|\frac{C(x(\cdot),\alpha_0+\delta\alpha) - C(x(\cdot),\alpha_0)}{\delta\alpha} - C_\alpha(x(\cdot),\alpha_0)\right|^2 = 0,$$

and for some constant K_1 and α in a neighborhood of α_0,

$$E_{\alpha_0}C^2(x(\cdot),\alpha) \leq K_1, \qquad E_{\alpha_0}C_\alpha^2(x(\cdot),\alpha) \leq K_1.$$

In cases such as (1.4a), (1.4b) where C does not depend explicitly on α, (A3.5) can be dropped.

Let $a_2 = \sigma_2\sigma_2'$. The solutions to the stochastic differential equation (3.1) for parameter values other than α_0 are defined via a Girsanov transformation [6, Chap. IV.4], as follows. For $\alpha_0 + \delta\alpha = \alpha \in A_0$, define

$$\delta b_2(t) = [b_2(x(\cdot),t,\alpha_0+\delta\alpha) - b_2(x(\cdot),t,\alpha_0)].$$

Let P_{α_0} denote the measure induced on $C^r[0,T]$ by $x(\cdot)$ under the parameter α_0 and define the measure $P_{\alpha_0+\delta\alpha}$ by the Radon–Nikodym derivative

$$dP_{\alpha_0+\delta\alpha}/dP_{\alpha_0} = \exp\left[\int_0^T [\sigma_2^{-1}(x(\cdot),t)\delta b_2(t)]'\,dw_2(t) - \frac{1}{2}\int_0^T |\sigma_2^{-1}(x(\cdot),t)\delta b_2(t)|^2\,dt\right].$$

Under $P_{\alpha_0+\delta\alpha}$, the processes $w_1(\cdot)$ and $w_2(t,\alpha_0+\delta\alpha) \equiv w_2(t) - \int_0^t \sigma_2^{-1}(x(\cdot),s)\delta b_2(s)\,ds$ are mutually independent Wiener processes. Also, rewriting (3.1) yields

$$dx = \begin{pmatrix} b_1(x(\cdot),t) \\ b_2(x(\cdot),t,\alpha_0+\delta\alpha) \end{pmatrix} dt + \begin{pmatrix} \sigma_1(x(\cdot),t)dw_1(t) \\ \sigma_2(x(\cdot),t)dw_2(t,\alpha_0+\delta\alpha) \end{pmatrix}. \tag{3.1a}$$

Thus $x(\cdot)$, under $P_{\alpha_0+\delta\alpha}$, is the solution process corresponding to parameter value $\alpha_0 + \delta\alpha$. Also, (3.1a) has a weak sense unique solution, since (3.1) does by (A3.3) [6, Chap. 4].

Define $Z(\cdot,\alpha_0)$ by

$$
\begin{aligned}
Z(t,\alpha_0) &= \int_0^t [\sigma_2^{-1}(x(\cdot),s)b_{2,\alpha}(x(\cdot),s,\alpha_0)]'dw_2(s) \\
&= \int_0^t [b'_{2,\alpha}(x(\cdot),s,\alpha_0)a_2^{-1}(x(\cdot),s)][dx_2(s) - b_2(x(\cdot),s,\alpha_0)ds].
\end{aligned} \tag{3.2}
$$

THEOREM 3.1. *Assume that the model* (3.1) *and* (A3.1)–(A3.5) *hold. Then* $V(\cdot)$ *is differentiable at* α_0 *and the quantity* $Q(\alpha_0)$ *defined by*

$$Q(\alpha_0) = Z(T,\alpha_0)C(x(\cdot),\alpha_0) + C_\alpha(x(\cdot),\alpha_0) \tag{3.3}$$

is an unbiased estimator of $V_\alpha(\alpha_0)$; *i.e.*, $E_{\alpha_0}Q(\alpha_0) = V_\alpha(\alpha_0)$.

Remark. Suppose that $C(x(\cdot),\alpha)$ takes the form

$$C(x(\cdot),\alpha) = \int_0^T k(x(\cdot,\alpha),s,\alpha)\,ds, \tag{3.4}$$

where k is nonanticipative and is bounded in the mean square sense. Then, since $E_{\alpha_0}k(x(\cdot),s,\alpha_0)[Z(T,\alpha_0) - Z(s,\alpha_0)] = 0$ for $s < T$, the quantity

$$\tilde{Q}(\alpha_0) = \int_0^T [k(x(\cdot),s,\alpha_0)Z(s,\alpha_0) + k_\alpha(x(\cdot),s,\alpha_0)]\,ds \tag{3.5}$$

is also an unbiased estimator of $V_\alpha(\alpha_0)$, and has a smaller variance than has $Q(\alpha_0)$. Numerical experiments indicate a variance reduction of a factor of $\frac{1}{2}$.

Suppose that an unbiased estimator $\overline{k}(s)$ of $k(x(\cdot),s,\alpha_0)$ with small variance is available. Then

$$\hat{Q}(\alpha_0) = \int_0^T \left[[k(x(\cdot),s,\alpha_0) - \overline{k}(s)]Z(s,\alpha_0) + k_\alpha(x(\cdot),s,\alpha_0)\right] ds \tag{3.5$'$}$$

will have smaller variance than $\tilde{Q}(\alpha_0)$. It is much easier to get excellent estimators for the $k(x(\cdot),\alpha,s)$ than for the derivatives and commonly the variances are negligible. The use of (3.5$'$) commonly reduces the variance by 20–50 percent.

Proof. Write

$$
\begin{aligned}
&[V(\alpha_0+\delta\alpha) - V(\alpha_0)]/\delta\alpha \\
&= [E_{\alpha_0+\delta\alpha}C(x(\cdot),\alpha_0+\delta\alpha) - E_{\alpha_0}C(x(\cdot),\alpha_0)]/\delta\alpha \\
&= E_{\alpha_0}[C(x(\cdot),\alpha_0+\delta\alpha)dP_{\alpha_0+\delta\alpha}/dP_{\alpha_0} - C(x(\cdot),\alpha_0)]/\delta\alpha \\
&= E_{\alpha_0}C(x(\cdot),\alpha_0)(dP_{\alpha_0+\delta\alpha}/dP_{\alpha_0} - 1)/\delta\alpha \\
&\quad + E_{\alpha_0}\left[\frac{C(x(\cdot),\alpha_0+\delta\alpha) - C(x(\cdot),\alpha_0)}{\delta\alpha}\right]\frac{dP_{\alpha_0+\delta\alpha}}{dP_{\alpha_0}}.
\end{aligned}
$$

Since $\lim_{\delta\alpha\to 0} E_{\alpha_0}|(dP_{\alpha_0+\delta\alpha})/dP_{\alpha_0} - 1|^2 = 0$, (A3.5) implies that the second right-hand term converges in mean to $E_{\alpha_0}C_\alpha(x(\cdot),\alpha_0)$. Also, it is readily shown that

$$\left(\frac{dP_{\alpha_0+\delta\alpha}}{dP_{\alpha_0}} - 1\right)/\delta\alpha \to Z(T,\alpha_0)$$

in mean square as $\delta\alpha \to 0$. The existence of $V_\alpha(\alpha_0)$ and representation (3.3) follow from these calculations. □

THE COMPUTATIONAL PROBLEM. Suppose that α is a vector with K components $(\alpha_1, \cdots, \alpha_K)$. Let $\alpha_0 = (\alpha_{0,1}, \cdots, \alpha_{0,K})$. To get an estimator of the gradient $V_\alpha(\alpha_0)$, we need to approximate the solution of (3.1) only once, and then compute the analogue of (3.2) for $b_{2,\alpha}$ replaced by b_{2,α_i} for each i. This computation is generally easier than getting a good approximation to either (3.1) or (2.3). The difference is particularly strong when (3.1) is a functional differential equation. Since $C(\cdot, \alpha)$ or $b(\cdot, \alpha)$ need not be x-differentiable here, cost functionals of the type (1.4b) and (1.7) can be included and so can systems of the type (1.6). Furthermore, we do not require that (3.1) have strong sense solutions, so that systems with discontinuous terms and that are defined by the Girsanov measure transformation method (such as (1.6)) can be used.

The idea of using stochastic methods for deterministic problems is sometimes attractive. To do this, we can simulate the deterministic system with a small amount of white noise added, and use (3.3). More will be said about this idea in §10. A similar procedure can be used if the stochastic system is not of the partitioned form (3.1). Simply add $\epsilon dw_3(\cdot)$ to the "dx_2 system," where $w_3(\cdot)$ is an R^{r_2}-valued Wiener process that is independent of $w_1(\cdot), w_2(\cdot)$, and ϵ is small, and use "corrections" analogous to that of §10.

Example (1.6)–(1.7). Here, $\alpha = (\alpha_1, \alpha_2)$, and $b_2(x, \alpha) = -u(x, \alpha)x^2 - \beta x^1 - \gamma \operatorname{sign} x^2$, where $u(\cdot)$ is defined by (1.8). For each $x, u(x, \alpha)$ is piecewise continuously differentiable in the components of α. For each α and $t > 0$, the probability is zero that $x^2(t) = 0$ and that the α-derivatives of $u(x(t), \alpha)$ and hence of $b_2(x(t), \alpha)$ do not exist. All the conditions of the theorem hold.

The jump-diffusion case. The method of Theorem 3.1 can be readily extended to the case where (3.1) is a jump-diffusion model. Let $D^r[0,T]$ denote the space of R^r-valued functions that are right continuous and have left-hand limits, and are continuous at T. The Skorokhod topology [7], [9] will be used on $D^r[0,T]$. Let $N(\cdot)$ be a Poisson measure [6, p. 42] with jump rate $\lambda < \infty$, and jump distribution $\Pi(\cdot)$ with compact support Λ. Let $q(\cdot)$ be a bounded measurable function on $D^r[0,T] \times [0,T] \times \Lambda$ with the property that $\lim_{s \uparrow t} \ q(\phi(\cdot), s, \gamma) = q(\phi(\cdot), t^-, \gamma)$ exists for all $\phi(\cdot) \in D^r[0,T]$. Let $q(\cdot)$ be nonanticipative in the sense used in (A3.1). Consider system (3.1) with the jump term $\int_\Lambda q(x(\cdot), t^-, \gamma) N(d\gamma dt)$ added [6; Chap. IV.9]:

$$dx = b(x(\cdot), t, \alpha)dt + \sigma(x(\cdot), t)dw + \int_\Lambda q(x(\cdot), t^-, \gamma)N(d\gamma dt). \tag{3.6}$$

Assume (A3.1)–(A3.5) with $C^r[0,T]$ replaced by $D^r[0,T]$. Then Theorems 3.1 and 3.2 continue to hold and the proof is unchanged.

The second derivative. The procedure of Theorem 3.1 can be repeated so that unbiased estimators of derivatives (or partial derivatives) of all orders can be obtained, if b_2 and C are smooth enough in α. For simplicity in the development, we let α be real valued. The general case should be clear.

THEOREM 3.2. *Assume* (A3.1)–(A3.5) *and the assumptions above on the jump terms. Let C and b_2 have second derivatives in the sense implied by* (A3.4) *and* (A3.5), *and let C be bounded. Define*

$$\begin{aligned} Q_2(\alpha_0) = &\left[Z^2(T, \alpha_0) - \int_0^T \left| \sigma_2^{-1}(x(\cdot), t) b_{2,\alpha}(x(\cdot), t, \alpha_0) \right|^2 dt \right] C(x(\cdot), \alpha_0) \\ &+ Z_\alpha(T, \alpha_0) C(x(\cdot), \alpha_0) + 2Z(T, \alpha_0) C_\alpha(x(\cdot), \alpha_0) + C_{\alpha\alpha}(x(\cdot), \alpha_0). \end{aligned} \tag{3.7}$$

Then

$$V_{\alpha\alpha}(\alpha_0) = E_{\alpha_0} Q_2(\alpha_0). \tag{3.8}$$

Proof. The proof is similar to that of Theorem 3.1, and only a few remarks will be made. For simplicity, suppose that C does not depend explicitly on α, and that we work with the diffusion case. For arbitrary $\alpha = \alpha_0 + \delta\alpha$, the process $Z(\cdot,\alpha)$ depends on $x(\cdot)$ and $w_2(\cdot,\alpha)$ and it is convenient to write this dependence explicitly by rewriting $Z(t,\alpha)$ in the form

$$\begin{aligned} &Z(t, \alpha_0 + \delta\alpha, w_2(\cdot, \alpha_0 + \delta\alpha), x(\cdot)) \\ &\quad = \int_0^t [\sigma_2^{-1}(x(\cdot), s) b_{2,\alpha}(x(\cdot), s, \alpha_0 + \delta\alpha)]'[dw_2(s) - \sigma_2^{-1}(x(\cdot), s)\delta b_2(s) ds], \end{aligned}$$

where $\delta b_2(s)$ is defined above Theorem 3.1.

Thus,

$$\begin{aligned} &(3.9) \\ &E_{\alpha_0+\delta\alpha} Z(T, \alpha_0 + \delta\alpha, w_2(\cdot, \alpha_0 + \delta\alpha), x(\cdot)) C(x(\cdot)) = E_{\alpha_0+\delta\alpha} Z(T, \alpha_0 + \delta\alpha, w_2(\cdot), x(\cdot)) \\ &\cdot C(x(\cdot)) - E_{\alpha_0+\delta\alpha} \int_0^T [\sigma_2^{-1}(x(\cdot), s) b_{2,\alpha}(x(\cdot), s, \alpha_0 + \delta\alpha)]' \sigma_2^{-1}(x(\cdot), s) \delta b_2(s)\, ds \cdot C(x(\cdot)). \end{aligned}$$

Dividing the second term on the right side of (3.9) by $\delta\alpha$ and taking $\delta\alpha \to 0$ yields

$$-E_{\alpha_0} \int_0^T |\sigma_2^{-1}(x(\cdot), s) b_{2,\alpha}(x(\cdot), s, \alpha_0)|^2\, ds \cdot C(x(\cdot)).$$

This is the second part of the first term of (3.7). The derivation of the other terms follows the lines of Theorem 3.1 and is omitted. □

4. A discrete-time approximation. Since we cannot obtain the paths $x(\cdot)$ or $w_2(\cdot)$ or the values of (3.3) exactly, some sort of approximation needs to be used. From a conceptual perspective, perhaps the simplest method is to use the following discrete-time form (with time-discretization parameter Δ):

$$x_{n+1}^\Delta = x_n^\Delta + b(x^\Delta(\cdot), n\Delta, \alpha_0)\Delta + \sigma(x^\Delta(\cdot), n\Delta)\delta w(n\Delta), \tag{4.1}$$

where $\delta w(n\Delta) = w(n\Delta + \Delta) - w(n\Delta)$. Let $x^\Delta(\cdot)$ be the piecewise constant interpolation of $\{x_n^\Delta\}$, defined by $x^\Delta(t) = x_n^\Delta$ on $[n\Delta, n\Delta + \Delta)$. Let $(x_{n,1}^\Delta, x_{n,2}^\Delta)$ denote the splitting of the components of x_n^Δ analogous to that in (3.1).

Define $Z^\Delta(\cdot, \alpha_0)$ to be the piecewise constant function with values

$$\begin{aligned} Z^\Delta(n\Delta, \alpha_0) &= \sum_{i=0}^{n-1} [\sigma_2^{-1}(x^\Delta(\cdot), i\Delta) b_{2,\alpha}(x^\Delta(\cdot), n\Delta, \alpha_0)]' \delta w_2(i\Delta) \\ (4.2) \qquad & \\ &= \sum_{i=0}^{n-1} [b_{2,\alpha}'(x^\Delta(\cdot), i\Delta, \alpha_0) a_2^{-1}(x^\Delta(\cdot), n\Delta)][\delta x_{i,2}^\Delta - b_2(x^\Delta(\cdot), i\Delta, \alpha)\Delta] \end{aligned}$$

on $[n\Delta, n\Delta + \Delta)$, where $\delta x_{i,2}^{\Delta} = x_{i+1,2}^{\Delta} - x_{i,2}^{\Delta}$. Let T be the integral multiple of Δ, and let the interpolations be left-continuous at T. By Theorem 4.1, the quantity

$$Q^{\Delta}(\alpha_0) = Z^{\Delta}(T, \alpha_0)C(x^{\Delta}(\cdot), \alpha_0) + C_{\alpha}(x^{\Delta}(\cdot), \alpha_0) \tag{4.3}$$

approximates (3.3). The analogue of (3.5) is

$$\sum_{i=0}^{T/\Delta-1} \Delta[k(x^{\Delta}(\cdot), i\Delta, \alpha_0)Z^{\Delta}(i\Delta, \alpha_0) + k_{\alpha}(x^{\Delta}(\cdot), i\Delta, \alpha_0)]. \tag{4.4}$$

To prove the convergence of $E_{\alpha_0}Q^{\Delta}(\alpha_0)$ to $V_{\alpha}(\alpha_0)$, we need to use the theory of weak convergence [7]–[9]. We will use the path space $D^p[0,T]$ with the Skorokhod topology, for appropriate p. The conditions used below are more restrictive than necessary, but they allow us to get to the main point easily. Assume the following.

(A4.1) $C(\cdot, \alpha_0), C_{\alpha}(\cdot, \alpha_0), \sigma_1(\cdot, \cdot), \sigma_2(\cdot, \cdot), b(\cdot, \cdot, \alpha)$, $b_{\alpha}(\cdot, \cdot, \alpha_0)$, and $\sigma_2^{-1}(\cdot, \cdot)$ are continuous functions on $D^r[0,T] \times [0,T]$, almost everywhere with respect to P_{α_0}, the measure induced by $x(\cdot, \alpha_0) = x(\cdot)$.

(A4.2) $C(\cdot, \alpha_0)$ and $C_{\alpha}(\cdot, \alpha_0)$ are bounded uniformly in A_0, $b(\cdot, \cdot, \alpha)$ and $\sigma(\cdot, \cdot)$ are bounded on bounded sets in $D^r[0,T]$ and uniformly in $\alpha \in A_0$.

(A4.3) Assumptions (A3.4) and (A3.5) hold with $x^{\Delta}(\cdot)$ replacing $x(\cdot)$, for each $\Delta > 0$.

Some preparatory calulations. The following representations are needed to state Theorem 4.1. Let P_{α}^{Δ} denote the measure induced by $x^{\Delta}(\cdot)$ on $D^r[0,T]$. Define $a = \sigma\sigma', a_2 = \sigma_2\sigma_2'$. Define

$$\begin{aligned} N(\alpha) = \exp -\frac{1}{2\Delta} \sum_{i=0}^{T/\Delta-1} & (\delta x_{i,2}^{\Delta} - b_2(x^{\Delta}(\cdot), i\Delta, \alpha)\Delta)' \\ & \cdot a_2^{-1}(x^{\Delta}(\cdot), i\Delta)(\delta x_{i,2}^{\Delta} - b_2(x^{\Delta}(\cdot), i\Delta, \alpha)\Delta), \\ \delta b_2^{\Delta}(i\Delta) = b_2(x^{\Delta}(\cdot), i\Delta, & \alpha_0 + \delta\alpha) - b_2(x^{\Delta}(\cdot), i\Delta, \alpha_0). \end{aligned}$$

Then

$$\begin{aligned} \frac{dP_{\alpha_0+\delta\alpha}^{\Delta}}{dP_{\alpha_0}^{\Delta}} &= \frac{N(\alpha_0 + \delta\alpha)}{N(\alpha_0)} \\ &= \exp\left[\sum_{i=0}^{T/\Delta-1} \delta b_2^{\Delta}(i\Delta)' a_2^{-1}(x^{\Delta}(\cdot), i\Delta)\delta x_{i,2}^{\Delta}\right. \\ &\quad - \Delta \sum_{i=0}^{T/\Delta-1} \delta b_2^{\Delta}(i\Delta)' a_2^{-1}(x^{\Delta}(\cdot), i\Delta) b_2(x^{\Delta}(\cdot), i\Delta, \alpha_0) \\ &\quad \left. - \frac{\Delta}{2} \sum_{i=0}^{T/\Delta-1} \delta b_2^{\Delta}(i\Delta)' a_2^{-1}(x^{\Delta}(\cdot), i\Delta)\delta b_2^{\Delta}(i\Delta)\right]. \end{aligned} \tag{4.5}$$

Thus under (A4.3),

$$\left(\frac{dP_{\alpha_0+\delta\alpha}^{\Delta}}{dP_{\alpha_0}^{\Delta}} - 1\right)\frac{1}{\delta\alpha} \to Z^{\Delta}(T, \alpha_0) \tag{4.6}$$

in mean square as $\delta\alpha \to 0$. Let E_{α}^{Δ} denote the expectation under parameter α and define $V^{\Delta}(\alpha) = E_{\alpha}^{\Delta}C(x^{\Delta}(\cdot), \alpha)$. By the next theorem, $E_{\alpha_0}^{\Delta}Q^{\Delta}(\alpha_0) = V_{\alpha}^{\Delta}(\alpha_0)$. Let $\Rightarrow$ denote weak convergence in the Skorokhod topology.

THEOREM 4.1. *Assume* (A3.1)–(A3.5) *and* (A4.1)–(A4.2). *Then*

$$\left(Z^{\Delta}(\cdot,\alpha_0), C(x^{\Delta}(\cdot),\alpha_0), C_{\alpha}(x^{\Delta}(\cdot),\alpha_0), x^{\Delta}(\cdot)\right)$$

$$\Rightarrow \left(Z(\cdot,\alpha_0), C(x(\cdot),\alpha_0), C_{\alpha}(x(\cdot),\alpha_0), x(\cdot)\right).$$

Also,

$$E^{\Delta}_{\alpha_0} Q^{\Delta}(\alpha_0) \to V_{\alpha}(\alpha_0). \tag{4.7}$$

Under the additional condition (A4.3),

$$E^{\Delta}_{\alpha_0} Q^{\Delta}(\alpha_0) = V^{\Delta}_{\alpha}(\alpha_0). \tag{4.8}$$

Remark on tightness. Let $\tau \le T$ denote a stopping time. Then a sufficient condition for tightness of $\{x^{\Delta}(\cdot), \Delta > 0\}$ is Theorem 2.7b of [9]:

$$\lim_{\delta\to 0} \overline{\lim_{\Delta}} \sup_{\tau} E|x^{\Delta}(\tau+\delta) - x^{\Delta}(\tau)|^2 = 0. \tag{4.9}$$

If $b(\cdot,\cdot,\alpha_0)$ and $\sigma(\cdot,\cdot)$ were bounded, then (4.9) holds. Otherwise, we exploit the uniqueness in (A3.3) and the almost everywhere continuity in (A4.1), and use a truncation procedure. The proof is standard and is omitted.

Extensions. Higher-order schemes. Methods for solving (3.1) of "higher order" than the simple "Euler" scheme (4.1), (4.2), can be used [16]. It is not necessarily advantageous to use such methods, however. In applications, many samples of the estimates are taken, and any extra time spent computing a single estimate will be at the expense of the number of samples taken. While higher-order methods might have smaller bias, it is the number of samples taken that determines the variance of the estimators, and both the bias and variance are important. Usually, in the early part of a Monte Carlo optimization method, smaller variance is more important than smaller bias.

5. A Markov chain approximation.

The form (4.1) is a "discrete time" approximation to (3.1). In this and the next two sections, we develop another computational approximation for (3.1), (3.2), which is a "discrete space" model. The method is based on an interpolated Markov chain approximation to $x(\cdot)$, and is a form of a method that has been very useful for solving the Bellman equation for optimal stochastic control problems [9], [10]. The idea can be used on a wide variety of problems. Experiments show that the numerical properties are roughly comparable to those of the method of §4, but are sometimes preferable in the sense of having smaller biases or variances. The method can also be extended to the controlled variance case (§8). When the underlying system (3.1) is unstable, it often yields better results than the method in §2.1 owing to the ease of controlling the "state increments." The method will be discussed first for the case of a "pure diffusion," and later for the jump diffusion. The basic technique will be defined in this section and a simple analogue of Theorem 4.1 proved. The next section concerns the convergence of the *derivative* of the cost for the chain to that for the diffusion, a fact that is not obvious. This result implies a "robustness" of the derivative estimates to the model. Section 7 concerns the convergence of the analogue of $Z(\cdot, \alpha_0)$ for the chain, and sheds some light on preferable constructions of the chain.

Definition of the approximating chain. We first describe the approximation when b and σ depend only on the current value of the state. We use h to denote the approximation parameter. For simplicity, we let it be scalar valued. For any parameter value α, let $\{\xi_n^h, n < \infty\}$ be a Markov chain that is locally consistent with $x(\cdot)$, under α, in the following sense. Let $E_{\alpha,n}^h$ denote expectation given $\{\xi_i^h, i \leq n\}$ and set $\delta\xi_n^h = \xi_{n+1}^h - \xi_n^h$. The parameter value α will be either stated, implied by $E_{\alpha,n}^h$ or else be α_0 by default. Suppose that there is a function $\Delta t^h(\cdot)$ (an interpolation interval) such that for $\xi_n^h = x$,

$$\begin{gathered}
E_{\alpha,n}^h \delta\xi_n^h = b(x,\alpha)\Delta t^h(x) + o(h^2), \\
E_{\alpha,n}^h \left[\delta\xi_n^h - E_{\alpha,n}^h \delta\xi_n^h\right]\left[\delta\xi_n^h - E_{\alpha,n}^h \delta\xi_n^h\right]' = a(x)\Delta t^h(x) + o(h^2), \\
|\delta\xi_n^h| = O(h), \\
c_0 h^2 \leq \Delta t^h(x) \leq c_1 h^2, \qquad c_i > 0.
\end{gathered} \tag{5.1}$$

We also suppose that the derivatives with respect to α of the $o(h^2)$ terms are $o(h^2)$. Define $\Delta t_n^h = \Delta t^h(\xi_n^h)$ and define $t_n^h = \sum_{i=0}^{n-1} \Delta t_i^h$. Define the continuous parameter interpolation $\xi^h(\cdot)$ by $\xi^h(t) = \xi_n^h$ on $[t_n^h, t_{n+1}^h)$. Under the uniqueness condition (A3.3) and the continuity of b and $\sigma, \xi^h(\cdot) \Rightarrow x(\cdot)$ for each α [9], [10]. The chain is easily simulated and provides a useful approximation to (3.1), (3.2). Let $p^h(y|x,\alpha)$ denote the one-step transition probabilities.

The stochastic functional differential equation case. In (3.1), the b and σ were allowed to depend on the "past" of the state process, and this can be readily carried over to the chain model, although we do lose the Markov property. Let $\Delta t^h(\cdot,\cdot)$ be a real-valued nonanticipative function (in the sense of (A3.1)) on $D^r[0,T] \times [0,T]$ satisfying

$$c_o h^2 \leq \Delta t^h(\phi(\cdot), t) \leq c_1 h^2,$$

where $\phi(\cdot)$ is the canonical point of $D^r[0,T]$. Define $\Delta t_n^h = \Delta t^h(\xi^h(\cdot), t_n^h)$, and replace the right-hand sides of the first two lines of (5.1) by

$$\begin{gathered}
b(\xi^h(\cdot,\alpha), t_n^h, \alpha)\Delta t_n^h + o(h^2), \\
a(\xi^h(\cdot,\alpha), t_n^h)\Delta t_n^h + o(h^2).
\end{gathered} \tag{5.2}$$

Here, we write the transition probability at time n as $p^h(y|\xi_i^h, i \leq n, \alpha)$.

Construction of the chain. A one-sided difference model. There are many methods for simulating chains with properties (5.1) or (5.2). Some general methods are in [9, Chap. 6] or [10, §5], where approximations based on both finite-difference and finite-element approximations to the differential operator of $x(\cdot)$ are given. The references deal with the pure Markov case, but the idea is trivially altered to suit the "functional" case, and we do that here. We next describe a basic scheme that is a functional form of a model in [9, Chap. 6], [10, §5], and then discuss variations and alternatives. The construction is only one suggestion but it illustrates the basic idea.

Let R_h^r be an h-grid on R^r (the set of points h units apart in each coordinate direction), and let e_i be the unit vector in the ith coordinate direction. Set $f^+ = \max(0, f), f^- = \max(0, -f)$. Define $\sigma\sigma' = a = \{a_{ij}\}$. Suppose that the current state

$\xi_n^h = x$. Then we let x communicate to $x \pm e_i h, x \pm e_i h \pm e_j h, x \pm e_i h \mp e_j h$, and construct the transition probability as follows.

For $\phi(\cdot) \in D^r[0,T]$, define

$$q_h(\phi(\cdot),t,\alpha) = \sum_i a_{ii}(\phi(\cdot),t) - \sum_{\substack{i \neq j \\ i,j}} |a_{ij}(\phi(\cdot),t)|/2 + h \sum_i |b_i(\phi(\cdot),t,\alpha)|, \tag{5.3}$$

and suppose that

$$a_{ii}(\phi(\cdot),t) - \sum_{j:j\neq i} |a_{ij}(\phi(\cdot),t)| \geq 0, \quad \text{for all} \quad t, \phi(\cdot), i. \tag{5.4}$$

Of course, $\phi(\cdot)$ represents the canonical path of $\xi^h(\cdot)$ for any α. Set

$$\Delta t^h(\phi(\cdot),t) = h^2/[\sum_i a_{ii}(\phi(\cdot),t) - \sum_{\substack{i \neq j \\ i,j}} |a_{ij}(\phi(\cdot),t)|/2]. \tag{5.5}$$

The Δt^h in [9] and [10] differs by $O(h^3)$ from that used here. Define

$$\begin{aligned} &p^h(x \pm e_i h | \xi_k^h, k < n, \xi_n^h = x, \alpha) \\ &\quad = [a_{ii}(\xi^h(\cdot),t_n^h)/2 - \sum_{j:j\neq i} |a_{ij}(\xi^h(\cdot),t_n^h)|/2 + h b_i^{\pm}(\xi^h(\cdot),t_n^h,\alpha)]/q_h(\xi^h(\cdot),t_n^h,\alpha), \end{aligned}$$

$$\begin{aligned} &p^h(x + e_i h + e_j h | \xi_k^h, k < n, \xi_n^h = x, \alpha) \\ &\quad = p^h(x - e_i h - e_j h | \xi_k^h, k < n, \xi_n^h = x, \alpha) \\ &\quad = a_{ij}^+(\xi^h(\cdot),t_n^h)/2q_h(\xi^h(\cdot),t_n^h,\alpha), \end{aligned} \tag{5.6}$$

$$\begin{aligned} p^h(x + e_i h - e_j h | \xi_i^h, i < n, \xi_n^h = x, \alpha) &= p^h(x - e_i h + e_j h | \xi_i^h, i < n, \xi_n^h = x, \alpha) \\ &= a_{ij}^-(\xi^h(\cdot),t_n^h)/2q_h(\xi^h(\cdot),t_n^h,\alpha). \end{aligned}$$

The process constructed from (5.6) satisfies (5.2) with Δt^h being given by (5.5) [9], [10]. Despite the formidable appearance of (5.6), it requires no more CPU time to simulate $\{\xi_n^h\}$ than the $\{x_n^\Delta\}$ defined in §4 does.

A central difference model. For $i = r_1 + 1, \cdots, r$ replace the $hb_i^{\pm}$ in the first equation of (5.6) by

$$\pm hb_i(\xi^h(\cdot),t_n^h,\alpha)/2, \tag{5.7}$$

and assume that h is small enough so that the numerators are nonnegative. Redefine q_h by dropping the $b_i, i > r_1$. (The $b_i, i \leq r_1$, do not depend on α.)

Discussion. State spaces such as R_h^r (or subsets of it) were used in [9], [10], since these references were concerned with a numerical solution to the Bellman equation for an optimal cost, and some "regularity" properties of the grid were essential for the programming. In the present problem, we are simply doing a simulation, and the scheme that is used can actually be changed at each step if we wish. There is considerably more flexibility, as long as (5.1) or (5.2) hold, as appropriate. For example, the stepsize can depend on the current value of x. To do this, we use c_i and an $h_o(\cdot)$ such that $c_3 h \le h_o(x) \le c_4 h$, and use $h_o(x)$ in lieu of h. In fact, it is often necessary to do such a "scaling" to get good numerical properties when the values of the dynamical terms vary considerably with x, particularly if the system is not stable. In addition, we need not use a grid that is aligned with the coordinate axes. The "local orientation" of the state transitions can depend on the current state. By aligning them with the "principle directions" of the matrix $\{a_{ij}\}$, the condition (5.4) can be eliminated.

The transition probabilities in (5.6) are no harder to calculate than are the coefficients in the discrete-time form (4.1), even for the functional dependence case. For a typical "functional" example, refer to the case discussed in §3, where the system is (3.1′), (3.1″). Then $\Delta t_n^h = \Delta t^h(\xi_n^h), a(\xi^h(\cdot), t_n^h) = a(\xi_n^h)$, and $b(\xi^h(\cdot), t_n^h, \alpha_0) = b(\xi_n^h, u_n^h)$ where $u_n^h = \sum_i^q u_{n,i}^h$ is the actual control action at step n. Let $\alpha_0 = (\beta_i, \gamma_i)$. Then letting $\xi^{h,1}$ denote the "position" component of ξ^h, we have

$$u_{n,i}^h = \sum_{k=0}^{n-1} e^{-\gamma_i (t_n^h - t_k^h)} \beta_i \xi_k^{h,1} \Delta t_k^h,$$

$$u_{n+1,i}^h = e^{-\gamma_i \Delta t_n^h} u_{n,i}^h + \beta_i \xi_n^{h,1} \Delta t^h(\xi_n^h).$$

Notation. Define

$$\delta W_n^{h,2} = \sigma_2^{-1}(\xi^h(\cdot), t_n^h)[\delta \xi_n^{h,2} - E_n^h \delta \xi_n^{h,2}],$$

and let $\tilde{Z}^h(\cdot, \alpha_0)$ be the piecewise constant process that takes the values on $[t_n^h, t_{n+1}^h)$:

$$\sum_{i=0}^{n-1} [\sigma_2^{-1}(\xi^h(\cdot), t_i^h) b_{2,\alpha}(\xi^h(\cdot), t_i^h, \alpha)]' \delta W_i^{h,2}. \tag{5.8}$$

For t in this interval, we can also write it in the form

$$\sum_{i:t_i^h \le t} b'_{2,\alpha}(\xi^h(\cdot), t_i^h, \alpha_0) a_2^{-1}(\xi^h(\cdot), t_i^h)[\delta \xi_i^{h,2} - b_2(\xi^h(\cdot), t_i^h \alpha_0) \Delta t_i^h] + \delta_h(t),$$

where $\delta_h(t) = O(ht)$. Note that $E_{\alpha_0,n}^h \delta W_n^{h,2} = 0$ and $E_{\alpha_0,n}^h \delta W_n^{h,2} (\delta W_n^{h,2})' = \Delta t_n^h I$. Hence $\delta W_n^{h,2}$ behaves "locally" like an increment to a Wiener process. Clearly, $\tilde{Z}^h(\cdot, \alpha_0)$ is an approximation to $Z(\cdot, \alpha_0)$, at least formally. Define

$$V^h(\alpha) = E_\alpha^h C(\xi^h(\cdot), \alpha),$$

$$\tilde{Q}^h(\alpha_0) = \tilde{Z}^h(T, \alpha_0) C(\xi^h(\cdot), \alpha_0) + C_\alpha(\xi^h(\cdot), \alpha_0).$$

The next theorem justifies the use of $\tilde{Q}^h(\alpha_0)$ as an asymptotically unbiased estimator of $V_\alpha(\alpha_0)$.

THEOREM 5.1. *Assume* (A3.1)–(A3.5), (A4.1)–(A4.2) *and let the chains satisfy the consistency conditions* (5.1) *or* (5.2). *Then, as* $h \to 0$,

$$\left(\tilde{Z}^h(\cdot, \alpha_0), C(\xi^h(\cdot), \alpha_0), C_\alpha(\xi^h(\cdot), \alpha_0), \xi^h(\cdot)\right)$$

$$\Rightarrow \Big(Z(\cdot,\alpha_0), C(x(\cdot),\alpha_0), C_\alpha(x(\cdot),\alpha_0), x(\cdot)\Big).$$

Also, as $h \to 0$

$$E^h_{\alpha_0}\tilde{Q}^h(\alpha_0) \to V_\alpha(\alpha_0).$$

For $\alpha \in A_0, \xi^h(\cdot) \Rightarrow x(\cdot)$ and the convergence is uniform in the sense that $V^h(\alpha) \to V(\alpha)$ uniformly in A_0.

Remark. The relationship between $E^h_{\alpha_0}\tilde{Q}^h(\alpha_0)$ and $V^h_\alpha(\alpha_0)$ will be made clear in the next section. The proof of the theorem is similar to that of Theorem 7.1, although Theorem 7.1 has some additional complications. In order not to duplicate the details, we omit the proof here, and refer the reader to §7. If the cost function has the form (3.4), then the obvious analogue of (3.5) is preferable.

6. Further properties of the chain approximation. In this section, we obtain an unbiased estimator of the derivative of the cost function for the chain $\{\xi^h_n\}$ at α_0 (Theorem 6.1) and show that its mean value converges to $V_\alpha(\alpha_0)$ as $h \to 0$ (Theorem 6.2). We will need the following assumptions. In fact, (A6.1) is a "typical" situation. In particular, it holds for the specific chains constructed in §5 if b_2 is α-differentiable.

(A6.1) There is $K_0 > 0$ such that for any $(n, y, \xi^h_i, i \le n)$ either $p^h(y|\xi^h_i, i \le n, \alpha)$ is zero for all $\alpha \in A_0$ or else it is $\ge K_0$ for all $\alpha \in A_0$. There is a bounded α-derivative p^h_α at α_0 in that

$$p^h(y|\xi^h_i, i \le n, \alpha_0 + \delta\alpha) = p^h(y|\xi^h_i, i \le n, \alpha_0) + \delta\alpha p^h_\alpha(y|\xi^h_i, i \le n, \alpha_0) + o(\delta\alpha),$$

where $o(\delta\alpha)$ is uniform in $y, \{\xi^h_i\}, n$.

(A6.2) Assumption (A3.5) holds for $\xi^h(\cdot)$ replacing $x(\cdot)$. The $C(\cdot,\cdot)$ and $C_\alpha(\cdot,\cdot)$ are bounded.

Define

$$v^h_i = p^h_\alpha(\xi^h_{i+1}|\xi^h_j, j \le i, \alpha_0)/p^h(\xi^h_{i+1}|\xi^h_j, j \le i, \alpha_0)$$

and define the piecewise constant process $Z^h(\cdot,\alpha_0)$ by $Z^h(t,\alpha_0) = \sum_{i=0}^{n-1} v^h_i$ on $[t^h_n, t^h_{n+1})$. Define

$$\begin{aligned} V^h(\alpha) &= E^h_\alpha C(\xi^h(\cdot),\alpha),\\ Q^h(\alpha_0) &= Z^h(T,\alpha_0)C(\xi^h(\cdot),\alpha_0) + C_\alpha(\xi^h(\cdot),\alpha_0),\\ N_h &= \min\{n : t^h_n \ge T\}. \end{aligned}$$

Note the N_h depends on the path, but not on the value of α directly. The asymptotic properties of $Z^h(\cdot,\alpha_0)$ will be developed in the next section.

THEOREM 6.1. *Assume* (A6.1) *and* (A6.2). *Then*

$$V^h_\alpha(\alpha_0) = E^h_{\alpha_0}Q^h(\alpha_0).$$

Proof. Let P^h_α denote the measure induced by the $\{\xi^h_i(\alpha), i \le N_h\}$. Then, by (A6.1), the P^h_α are mutually absolutely continuous and on any path $\{\xi^h_i, i \le N_h\}$,

$$\frac{dP^h_{\alpha_0+\delta\alpha}}{dP^h_{\alpha_0}} = \prod_{i=0}^{N_h-1} \frac{p^h(\xi^h_{i+1}|\xi^h_j, j \le i, \alpha_0 + \delta\alpha)}{p^h(\xi^h_{i+1}|\xi^h_j, j \le i, \alpha_0)}.$$

By (A6.1), we have

$$\frac{dP^h_{\alpha_0+\delta\alpha}}{dP^h_{\alpha_0}} = \prod_{i=0}^{N_h-1} (1 + v_i^h \delta\alpha + o(\delta\alpha)).$$

As in the proof of Theorem 3.1, we have

$$\begin{aligned}&\frac{V^h(\alpha_0+\delta\alpha) - V^h(\alpha_0)}{\delta\alpha}\\ &= E^h_{\alpha_0}\left[C(\xi^h(\cdot),\alpha_0)\left(\frac{dP^h_{\alpha_0+\delta\alpha}}{dP^h_{\alpha_0}} - 1\right)\Big/\delta\alpha + C_\alpha(\xi^h(\cdot),\alpha_0)\right] + \epsilon_1(\delta\alpha),\end{aligned}$$

where $\epsilon_1(\delta\alpha) \to 0$ as $\delta\alpha \to 0$. By the above calculations,

$$\left(\frac{dP^h_{\alpha_0+\delta\alpha}}{dP^h_{\alpha_0}} - 1\right)\Big/\delta\alpha \to Z^h(T,\alpha_0)$$

in the mean square sense as $\delta\alpha \to 0$, and the proof is concluded. □

In the next theorem we show that $Q^h(\alpha_0)$ is a good estimator for $V_\alpha(\alpha_0)$ for small enough h. We will require the following definition and condition. Define

$$v_i^h(\alpha_0,\delta\alpha) = \frac{p^h(\xi^h_{i+1}|\xi^h_j, j \le i, \alpha_0+\delta\alpha)}{p^h(\xi^h_{i+1}|\xi^h_j, j \le i, \alpha_0)} - 1.$$

(A6.3) $v_i^h(\alpha_0,\delta\alpha) = O(\delta\alpha h)$, where $O(\cdot)$ is uniform in all other variables.

Remark on A6.3. Condition (A6.3) seems to be a "typical" case in applications. See, for example, the model (5.6) or (5.7), where if $b(x(\cdot),t,\alpha)$ is Lipschitz continuous in α, uniformly in the other variables, then (A6.3) holds. It follows from (A6.3) and (A6.1) that $p^h_\alpha(y|\xi^h_i, i \le n, \alpha_0) = O(h)$ uniformly in all the variables. Note that (A6.3) implies that $v_i^h = O(h)$.

THEOREM 6.2. *Assume* (A3.1)–(A3.5), (A4.1)–(A4.2), *and* (A6.1), (A6.3). *Let* $\{\xi^h_n\}$ *satisfy the consistency condition* (5.1) *or* (5.2), *for each* α. *Then*

$$E^h_{\alpha_0} Q^h(\alpha_0) \to V_\alpha(\alpha_0).$$

Proof. For simplicity, we will suppose that $C(x(\cdot),\alpha)$ does not depend on α. We will first show that

$$(dP^h_{\alpha_0+\delta\alpha}/dP^h_{\alpha_0} - 1)/\delta\alpha$$

is approximated in an appropriate sense by $Z^h(T,\alpha_0)$, uniformly in h. Since $\sum_y p^h(y|\xi^h_i, i \le n, \alpha_0+\delta\alpha) \equiv 1$, we have

(6.1)

$$E^h_{\alpha_0,n} v_n^h(\alpha_0,\delta\alpha) + 1 = \sum_y \left\{\left[\frac{p^h(y|\xi^h_i, i \le n, \alpha_0+\delta\alpha)}{p^h(y|\xi^h_i, i \le n, \alpha_0)}\right] p^h(y|\xi^h_i, i \le n, \alpha_0)\right\} = 1.$$

Thus, $\sum_{i=0}^n v_i^h(\alpha_0,\delta\alpha)/\delta\alpha$ is a martingale sequence (under $P^h_{\alpha_0}$) and the variance is $nO(h^2)$, by (A6.3). By a similar calculation, we get that $E^h_{\alpha_0,n} v_n^h = 0$. Thus,

$\{\sum_{i=0}^{n} v_i^h\}$ is a martingale sequence with respect to the measure $P_{\alpha_0}^h$. By (A6.3), it has variance $O(h^2)n$. Since on the interval of interest, we have $n \leq N_h = O(T/h^2)$ by (5.1) or (5.2), the variances are bounded. Note also that $E_{\alpha_0,n}^h v_i^h(\alpha_0, \delta\alpha) v_j^h = 0$ for $i, j \geq n$ and $i \neq j$.

For an integer M, define the set $B(M, \delta\alpha) = \{|\sum_{i=0}^{N_h - 1} v_i^h(\alpha_0, \delta\alpha)| \geq M\delta\alpha\}$. Then, using the above estimates, we can show that

$$E_{\alpha_0}^h \left[\left[\prod_{i=0}^{N_h-1} (1 + v_i^h(\alpha_0, \delta\alpha)) - 1 \right] \Big/ \delta\alpha - \sum_{i=0}^{N_h-1} v_i^h \right] I_{B(M,\delta\alpha)} \to 0$$

as $M \to \infty$, uniformly in h and $\delta\alpha$. Thus

$$\frac{V^h(\alpha_0 + \delta\alpha) - V^h(\alpha_0)}{\delta\alpha}$$
$$= E_{\alpha_0}^h C(\xi^h(\cdot), \alpha_0) \left[\prod_{i=0}^{N_h-1} (1 + v_i^h(\alpha_0, \delta\alpha)) - 1 \right] (1 - I_{B(M,\delta\alpha)}) \Big/ \delta\alpha$$

modulo an error that goes to zero uniformly in h and $\delta\alpha$ as $M \to \infty$.

For the path not in $B(M, \delta\alpha)$, we can use the "exponential" approximation to the product and the bound $O(h)$ on v_i^h and $v_i^h(\alpha_0, \delta\alpha)/\delta\alpha$ to get

$$\begin{aligned} \prod_{i=0}^{N_h-1} (1 + v_i^h(\alpha_0, \delta\alpha)) &= \exp \sum_{i=0}^{N_h-1} [v_i^h(\alpha_0, \delta\alpha) + O^2(h\delta\alpha)] \\ &= 1 + \sum_{i=0}^{N_h-1} v_i^h \delta\alpha + o(\delta\alpha), \end{aligned}$$

where $o(\delta\alpha)$ depends on M, but is uniform in h and ω. Using the above estimates and letting $M \to \infty$ yields

$$\frac{V^h(\alpha_0 + \delta\alpha) - V^h(\alpha_0)}{\delta\alpha} = E_{\alpha_0}^h Q^h(\alpha_0) + \epsilon(\delta\alpha, h),$$

where $\epsilon(\delta\alpha, h) \to 0$ as $\delta\alpha \to 0$ and $h \to 0$.

By Theorem 5.1, $V^h(\alpha_0 + \delta\alpha) \to V(\alpha_0 + \delta\alpha)$ for any $\alpha_0 + \delta\alpha \in A_0$, and the rate of convergence is uniform in $\delta\alpha$. The sequence $\{Q^h(\alpha_0)\}$ is tight. Let h_n index a weakly convergent subsequence of $\{Q^h(\alpha_0)\}$ and of $\{\xi^h(\cdot)\}$ under both parameters $\alpha_0 + \delta\alpha$ and α_0, with the limit of $\{Q^{h_n}(\alpha_0)\}$ denoted by Q. By the weak convergence,

$$\frac{V(\alpha_0 + \delta\alpha) - V(\alpha_0)}{\delta\alpha} = EQ + \epsilon_1(\delta\alpha),$$

where $\epsilon_1(\delta\alpha) = \lim_{h \to 0} \epsilon(\delta\alpha, h) \to 0$ as $\delta\alpha \to 0$. By Theorem 3.1, $V_\alpha(\alpha_0)$ exists. Thus $EQ = V_\alpha(\alpha_0)$. Hence $E_{\alpha_0} Q$ does not depend on the chosen subsequence, and $E_{\alpha_0}^h Q^h(\alpha_0) \to V_\alpha(\alpha_0)$. □

7. The limits of $\mathbf{Z^h}$ for the chain approximation. Theorem 6.2 proved that $Z^h(T, \alpha_0)$ can be used to get an asymptoptically consistent estimator of $V_\alpha(\alpha_0)$. It is of interest to know whether $Z^h(T, \alpha_0) \Rightarrow Z(T, \alpha_0)$ as $h \to 0$. We will see in Theorem 7.1 that this is not always the case, and the proof will yield insight into the preferable numerical methods. Essentially, when the convergence is not the case, then the limit of $Z^h(T, \alpha_0)$ equals $Z(T, \alpha_0)$ plus an "error," where the error increases the variance of the estimator, but does not affect the mean. The "error" can be eliminated by an appropriate choice of the chain. Next, we present an example. Define $m_h(t) = \max\{i : t_i^h \le t\}$.

Example. Let σ and b depend only on the current state (i.e., no functional dependence) and suppose that $a_{ij}(x) = 0$ for $i \neq j$ and $a_{ii}(x) > 0$ all x and all i. The same result will hold if $a_{ii}(x) > 0$ for only some of the i and each x. We use the "central difference" method (5.7) on the grid R_h^r. The "one-sided difference" method of (5.6) yields the same results. Let $a_{ii}(x) - h|b_i(x, \alpha_0)| \ge 0$, all x and i. We have

$$p^h(x \pm e_i h | x, \alpha_0) = [a_{ii}(x) \pm hb_i(x, \alpha_0)] \Big/ \Big(2 \sum_j a_{jj}(x)\Big),$$

$$\Delta t^h(x) = h^2 / \sum_i a_{ii}(x).$$

We must have $\delta\xi_n^h = \pm he_i$, for some i. Given that value i, we have the following representation:

$$\begin{aligned} v_n^h &= \pm hb_{i,\alpha}(\xi_n^h, \alpha_0)/[a_{ii}(\xi_n^h) \pm hb_i(\xi_n^h, \alpha_0)] \\ &= \frac{\pm hb_{i,\alpha}(\xi_n^h, \alpha_0)}{a_{ii}(\xi_n^h)} \left[1 \mp \frac{hb_i(\xi_n^h, \alpha_0)}{a_{ii}(\xi_n^h)}\right] + o(h^2). \end{aligned}$$

We can rewrite the principal part of the above expression as

$$[b_\alpha'(\xi_n^h, \alpha_0) a^{-1}(\xi_n^h)][\delta\xi_n^h - \delta Y_n^h] + o(h^2), \tag{7.1}$$

where δY_n^h is a vector with jth component

$$h^2 b_j(\xi_n^h, \alpha_0) I_{\{\delta\xi_n^h = \pm e_j h\}} / a_{jj}(\xi_n^h).$$

From the transition probability above, we can calculate the probability, given the past, that $\delta\xi_n^h = \pm he_i$. Using this, the conditional expectation of the jth component, given the "past," is calculated to be

$$E_{\alpha_0,n}^h \delta Y_n^{h,j} = h^2 b_j(\xi_n^h, \alpha_0) \Big/ \sum_k a_{kk}(\xi_n^h) + o(h^2) = \Delta t^h(\xi_n^h) b_j(\xi_n^h, \alpha_0) + o(h^2) = O(h^2).$$

It can be readily shown that

$$E\left|\sum_{i=0}^{n-1} (\delta Y_i^h - E_{\alpha_0,i}^h \delta Y_i^h)\right|^2 = nO(h^4).$$

Thus, we can rewrite the sum of the terms (7.1) over $[0, m_h(t))$ as

$$\begin{aligned} &\sum_{n=0}^{m^h(t)} [b_\alpha'(\xi_n^h, \alpha_0) a^{-1}(\xi_n^h)](\delta\xi_n^h - b(\xi_n^h, \alpha_0)\Delta t_n^h) \\ &\qquad + (\text{term that goes to zero as } h \to 0). \end{aligned} \tag{7.2}$$

It can be shown that the sum in (7.2) converges weakly to $Z(\cdot, \alpha_0)$ as $h \to 0$, under the conditions of Theorem 6.2. Thus, for this case, the conclusions of Theorem 5.1 hold for $Z^h(\cdot, \alpha_0)$ replacing $\tilde{Z}^h(\cdot, \alpha_0)$. (That is, the two are asymptotically equivalent.)

If $a(\cdot)$ is not diagonal and a transition function such as (5.6) (or with the modification (5.7)) used, then it follows from the proof of Theorem 7.1 that $Z^h(\cdot,\alpha_0) \to Z(\cdot,\alpha_0)+\hat{Z}(\cdot)$, where $\hat{Z}(\cdot)$ is not necessarily identically zero and $EC(x(\cdot),\alpha_0)\hat{Z}(T) = 0$. Then $\hat{Z}(\cdot)$ contributes nothing to the mean value of the estimate, but causes an increase in the variance of the estimator. The problem can be readily avoided by "rotating" the local coordinates such that the local transitions are essentially along the principle directions of $a(\xi_n^h)$.

THEOREM 7.1. *Assume the conditions of Theorem* 6.2. *Then* $E[Z^h(T,\alpha_0)C(\xi^h(\cdot),\alpha_0) + C_\alpha(\xi^h(\cdot),\alpha_0)] \to V_\alpha(\alpha_0)$. *Let h index a weakly convergent subsequence of* $\{Z^h(\cdot,\alpha_0)\}$. *Then* $Z^h(\cdot,\alpha_0) \Rightarrow Z(\cdot,\alpha_0)+\hat{Z}(\cdot)$, *where* $Z(\cdot,\alpha_0)$ *and* $\hat{Z}(\cdot)$ *are orthogonal martingales and* $EC(x(\cdot),\alpha_0)\hat{Z}(T) = 0$.

Proof. As in Theorem 4.1, we need only work with the "truncated" b and σ. For this reason, as well as to simplify the development, we suppose that b and σ are bounded. Also, for simplicity, we drop the C_α term. To get the weak convergence, we use the martingale method (see, e.g., [8], [10, Thm. 4.5]), but the proof here will be self contained. First, tightness will be proved. Let $E^h_{\alpha_0,t}$ denote the expectation conditioned on the data up to interpolated time t or, equivalently, up to discrete time $m_h(t)$. For any real-valued $\delta > 0$ and τ being any stopping time (taking values $t_i^h, i \le N_h$), (5.1) or (5.2) imply that

(7.3)
$$\begin{aligned} E^h_{\alpha_0,\tau}|\xi^h(\tau+\delta)-\xi^h(\tau)|^2 &\le E^h_{\alpha_0\tau}\sum_{n=m_h(\tau)+1}^{m_h(\tau+\delta)}\{|\delta\xi_n^h - E^h_{\alpha_0,n}\delta\xi_n^h|^2 + |E^h_{\alpha_0,n}\delta\xi_n^h|^2\} + O(h) \\ &\le \text{ const } \cdot E^h_{\alpha_0,\tau}\left[\sum_{n=m_h(\tau)+1}^{m_h(\tau+\delta)}\Delta t^h(\xi^h(\cdot),t_n^h)\right] + O(h) \\ &= O(\delta)+O(h). \end{aligned}$$

Also,

$$(7.4)\quad E^h_{\alpha_0}|Z^h(\tau+\delta,\alpha_0)-Z^h(\tau,\alpha_0)|^2 = E^h_{\alpha_0}\left|\sum_{n=m_h(\tau)+1}^{m_h(\tau+\delta)} v_n^h\right|^2 = O(\delta)+O(h).$$

Then tightness of $\{Z^h(\cdot,\alpha_0),\xi^h(\cdot),h>0\}$ follows from (7.3), (7.4), and the criterion (4.9). Since the discontinuities in $\xi^h(\cdot)$ and $Z^h(\cdot,\alpha_0)$ are $O(h)$, the limit of any weakly convergent subsequence must have continuous paths with probability one.

We next obtain a local expansion of a test function that will be used to characterize the differential operator of the limits of the weakly convergent subsequences. Let $f(\cdot)$ be a smooth real-valued function on R^{r+1} with compact support and define $Z_n^h = \sum_{i=0}^{n-1} v_i^h$. Then, using the fact that $|\delta\xi_n^h| + |v_n^h| = O(h)$, we can write

(7.5)
$$\begin{aligned} E^h_{\alpha_0,n}f(\xi^h_{n+1},Z^h_{n+1}) - f(\xi_n^h,Z_n^h) &= E^h_{\alpha_0,n}[f_x'(\xi_n^h,Z_n^h)\delta\xi_n^h + f_z(\xi_n^h,Z_n^h)v_n^h] \\ &\quad + E^h_{\alpha_0,n}[\tfrac{1}{2}(\delta\xi_n^h)'f_{xx}(\xi_n^h,Z_n^h)\delta\xi_n^h \\ &\quad + \tfrac{1}{2}f_{zz}(\xi_n^h,Z_n^h)(v_n^h)^2 + (\delta\xi_n^h)'f_{zx}(\xi_n^h,Z_n^h)v_n^h] \\ &\quad + o(h^2). \end{aligned}$$

Note that by the consistency condition (5.1) or (5.2),

$$
\begin{aligned}
E^h_{\alpha_0,n} v^h_n \delta\xi^h_n &= \sum_y \left[\frac{(y-\xi^h_n)p^h_\alpha(y|\xi^h_i, i \le n, \alpha_0)}{p^h(y|\xi^h_i, i \le n, \alpha_0)} \right] p^h(y|\xi^h_i, i \le n, \alpha_0) \\
&= \sum_y (y-\xi^h_n)p^h_\alpha(y|\xi^h_i, i \le n, \alpha_0) = \frac{\partial}{\partial\alpha} E^h_{\alpha_0,n}\delta\xi^h_n \qquad (7.6)\\
&= \frac{\partial}{\partial\alpha} b(\xi^h(\cdot), t^h_n, \alpha_0)\Delta t^h(\xi^h(\cdot), t^h_n) + o(h^2).
\end{aligned}
$$

Define the bounded (uniformly in $h, n \le N_h$, since $\Delta t^h(t) \ge c_o h^2$ and $v^h_n = O(h)$) sequence $\overline{a}^h_n$ by

$$
\begin{aligned}
E^h_{\alpha_0,n}(v^h_n)^2 &= \sum_y \frac{p^h_\alpha(y|\xi^h_i, i \le n, \alpha_0)^2}{p^h(y|\xi^h_i, i \le n, \alpha_0)} \qquad (7.7)\\
&= \overline{a}^h_n \Delta t^h(\xi^h(\cdot), t^h_n).
\end{aligned}
$$

Define the piecewise constant function $\overline{A}^h(\cdot)$ by $\overline{A}^h(t) = \sum_{i=0}^{n-1} \overline{a}^h_i \cdot \Delta t^h(\xi^h(\cdot), t^h_n)$ on $[t^h_n, t^h_{n+1})$. The sequence $\{\overline{A}^h(\cdot), h > 0\}$ is easily shown to be tight and all the weak limits are continuous. In fact, the weak limits must be Lipschitz continuous, hence almost everywhere differentiable.

The martingale problem. For any integer q and any $t, s > 0$, let $t_i \le t \le t+s, i \le q$. Let $h(\cdot)$ be a real-valued bounded and continuous function of its arguments. For square matrices $F = \{F_{ij}\}$ and $G = \{G_{ij}\}$, recall the relationship $\sum_{i,j} F_{ij}G_{ji} =$ trace $F \cdot G$. By (7.5)–(7.7) and a truncated Taylor expansion, we can write

$$
\begin{aligned}
&E^h_{\alpha_0} h(\xi^h(t_i), Z^h(t_i, \alpha_0), \overline{A}^h(t_i), i \le q) \\
&\quad \times \Bigg[f(\xi^h(t+s), Z^h(t+s, \alpha_0)) - f(\xi^h(t), Z^h(t, \alpha_0)) \\
&\quad - \sum_{n=m_h(t)+1}^{m_h(t+s)} \bigg\{ f'_x(\xi^h_n, Z^h_n) b(\xi^h(\cdot), t^h_n, \alpha_0) + \frac{1}{2} \text{ trace } f_{xx}(\xi^h_n, Z^h_n) a(\xi^h(\cdot), t^h_n) \qquad (7.8)\\
&\qquad + \frac{1}{2} f_{zz}(\xi^h_n, Z^h_n)\overline{a}^h_n + b'_\alpha(\xi^h(\cdot), t^h_n, \alpha_0) f_{zx}(\xi^h_n, Z^h_n) \bigg\} \Delta t^h_n \Bigg] \\
&= E^h_{\alpha_0} \sum_{n=0}^{N_h-1} o(h^2) \to 0.
\end{aligned}
$$

Abusing notation, let h also index a weakly convergent subsequence of $\{\xi^h(\cdot), Z^h(\cdot, \alpha_0), \overline{A}^h(\cdot), h > 0\}$ with limit denoted by $\{(x(\cdot), Z(\cdot), \overline{A}(\cdot)\}$. Define $\overline{a}(\cdot)$ by $\overline{A}(t) = \int_0^t \overline{a}(s)ds$. Then using the weak convergence and the continuity conditions (A4.1), we have

$$
\begin{aligned}
(7.9) \quad Eh(x(t_i), Z(t_i), \overline{A}(t_i), i \le q) &\Bigg[f(x(t+s), Z(t+s)) - f(x(t), Z(t)) \\
&- \int_t^{t+s} du \bigg\{ f'_x(x(u), Z(u)) b(x(\cdot), u, \alpha_0)
\end{aligned}
$$

$$\left.\left.+\frac{1}{2}\text{trace } f_{xx}(x(u),Z(u))\cdot a(x(\cdot),u) +\frac{1}{2}f_{zz}(x(u),Z(u))\overline{a}(u)+b'_\alpha(x(\cdot),u,\alpha_0) \cdot f_{zx}(x(u),Z(u))\right\}\right] = 0.$$

The arbitariness of $q, h(\cdot), f(\cdot), t, s, \{t_i\}$ in (7.9) implies that $(x(\cdot), Z(\cdot))$ solves the martingale problem for the operator $A(\alpha_0)$ that is defined by the bracketed expression in (7.9), and with respect to the filtration $\mathcal{B}_t = \mathcal{B}(x(s), Z(s), \overline{A}(s), s \leq t)$. The form of the operator implies that the process $Z(\cdot)$ is a $\mathcal{B}_t$-martingale (since there is no first-order term f_z), and that $x(\cdot)$ satisfies (3.1) for some mutually independent $\mathcal{B}_t$-Wiener processes $w_1(\cdot), w_2(\cdot)$. (If $\sigma_1(\cdot)$ is degenerate, then we might have to augment the probability space by adding an "independent" Wiener process.)

We wish to show next that $Z(\cdot)$ can be represented as

$$Z(t) = Z(t,\alpha_0) + \hat{Z}(t) = \int_0^t [\sigma_2^{-1}(x(\cdot),u)b_{2,\alpha}(x(\cdot),u,\alpha_0)]'\,dw_2(u) + \hat{Z}(t), \tag{7.10}$$

where $\hat{Z}(\cdot)$ and $Z(\cdot,\alpha_0)$ are orthogonal stochastic integrals. Define $X(t) = \int_0^t \sigma_2(x(\cdot),u)\,dw_2(u)$, the martingale part of $x_2(\cdot)$. The mutual quadratic variation process (written as a row vector) is

$$\Big\langle X(\cdot), Z(\cdot)\Big\rangle(t) = \int_0^t b'_{2,\alpha}(x(\cdot),u,\alpha_0)\,du = \Big\langle X(\cdot), Z(\cdot,\alpha_0)\Big\rangle(t). \tag{7.11}$$

The first equality of (7.11) follows from the form of the operator $A(\alpha_0)$ or (equivalently) the representation (7.9), and the second follows from the definition of $Z(\cdot,\alpha_0)$. Also $< Z(\cdot) > (t) = \int_0^t \overline{a}(s)\,ds$. With the definition $\tilde{a}(t) = b'_{2,\alpha}(x(\cdot),t,\alpha_0)a_2^{-1}(x(\cdot),t)$ $b_{2,\alpha}$ $(x(\cdot),t,\alpha_0)$, we have

$$\Big\langle Z(\cdot,\alpha_0)\Big\rangle(t) = \int_0^t \tilde{a}(s)\,ds = \Big\langle Z(\cdot,\alpha_0), Z(\cdot)\Big\rangle(t). \tag{7.12}$$

The first equality follows from the definition of $Z(\cdot,\alpha_0)$, while the second follows from that definition and (7.11), if we note that $Z(\cdot,\alpha_0)$ can be defined as a stochastic integral with respect to the martingale $X(\cdot)$.

The equalities in (7.12) imply [12] the decomposition (7.10), where $< \hat{Z}(\cdot), Z(\cdot,\alpha_0) > (t) \equiv 0$. Finally, by Theorem 6.2,

$$EZ^h(T,\alpha_0)C(\xi^h(\cdot),\alpha_0) \to V_\alpha(\alpha_0) = E_{\alpha_0}Z(T,\alpha_0)C(x(\cdot),\alpha_0).$$

Hence $E_{\alpha_0}\hat{Z}(T)C(x(\cdot),\alpha_0) = 0$. □

Remark. Note that $< Z > (t) =< Z(\alpha_0) > (t)+ < \hat{Z} > (t)$. Hence, $\hat{Z}(t) = \int_0^t[\overline{a}(s) - \tilde{a}(s)]\,ds$. With the transition probabilities given, it is often possible to compute $\overline{a}$ and $\tilde{a}$, hence to evaluate the increase in variance of the estimator.

The jump-diffusion case. The discrete-time approximation of §4, and the Markov chain approximation of §5 can be readily extended to the jump-diffusion case (3.6) when neither the jump rate nor the jump distribution depend on α. We comment only on the Markov chain case. A Markov chain approximation to a jump diffusion was developed in [11]. The reference [11] was concerned with solving the Bellman or related equations for mean values of functionals of the process and this required that the state space for the Markov chain approximation be "nice," i.e., the jumps can take the state to only a finite set of points in the state space. Since we are not interested in solving for functionals of the chain explicitly, we have considerably more flexibility here than in [11].

Let γ denote the canonical "jump" of the Poisson measure $N(\cdot)$, and let the random variable γ_h be an approximation to γ with only a finite number of values and such that

$$\sup_{\phi(\cdot),t} |q(\phi(\cdot),t,\gamma_h) - q(\phi(\cdot),t,\gamma)| \xrightarrow{h} 0,$$

where $\phi(\cdot)$ is the canonical variable of $D^r[0,T]$. Let $\{\xi_n^h\}$ denote the approximating chain under the parameter value α. Let $\xi_n^h = x$ and $t_n^h = t$. To be consistent with (3.6), the probability of no jump at step n can be either $1-\lambda\Delta t^h(\xi^h(\cdot),t)$ (or $e^{-\lambda\Delta t^h(\xi^h(\cdot),t)}$, if we wish). In that case, let $\xi_{n+1}^h - x = O(h)$ and satisfy (5.1) or (5.2), according to the case. The probability of a jump is $\lambda\Delta t^h(\xi^h(\cdot),t)$ or $1-e^{-\lambda\Delta t^h(\xi^h(\cdot),t)}$. Then with this probability, we choose ξ_{n+1}^h to satisfy

$$\xi_{n+1}^h - x = q(\xi^h(\cdot),t,\tilde{\gamma}_h),$$

where $\tilde{\gamma}_h$ has the distribution of γ_h and is independent of $\xi_i^h, i \le n$.

All the previous theorems of §§5–7 continue to hold, and the proofs do not change. We note, in particular, that if there is a "jump" at stage n, then $p^h(\xi_{n+1}^h = y|\xi_i^h, i \le n/\alpha)$ does not depend on α. Hence $v_n^h = 0$ for those values of n at which there is a jump.

8. Variance depending on the control parameter. Up to this point, the variance was not allowed to depend on the control parameter α. The basic reason was that the measures P_α would not then be mutually absolutely continuous, so that the basic ideas of Theorem 3.1 could not be performed. For example, let P_α denote the measure induced on $D[0,T]$ by $\alpha w(\cdot), \alpha \ne 0$, where $w(\cdot)$ is a standard Wiener process. Then the P_α are not mutually absolutely continuous. Despite the lack of mutual absolute continuity, there is one interesting class for which the Markov chain method of §5 can be used to show that $V_\alpha(\alpha_0)$ exists and also to yield an estimator of it whose bias goes to zero as $h \to 0$. The direct discrete-time approximation of §4 does not work for this case.

The method is related to the time change argument that is used to construct a solution to a one-dimensional diffusion. It seems to work only when the state process is real valued. We will work with the process

$$dx = b(x,\alpha)dt + \sigma(x,\alpha)dw, \tag{8.1}$$

and cost

$$C(x(\cdot),\alpha) = \int_0^T k(x(s),\alpha)ds + g(x(T),\alpha).$$

The results of Theorem 8.1 below also hold if b, σ, and k are nonanticipative functionals of the state. However, in view of an already complicated notation, we let them be functions of only the present value of the state. We will need the following assumption.

(A8.1) $\sigma(x,\cdot)$ is continuously differentiable in α in A_0 in that there is a bounded continuous function $\sigma_\alpha(\cdot,\cdot)$ such that

$$\left|\frac{\sigma(x,\alpha+\delta\alpha)-\sigma(x,\alpha)}{\delta\alpha}-\sigma_\alpha(x,\alpha)\right|\to 0$$

as $\delta\alpha\to 0$, uniformly in $R\times A_0$, and similarly for $b(\cdot,\cdot), k(\cdot,\cdot)$, and $g(\cdot,\cdot)$. Also, $k(\cdot,\cdot), b(\cdot,\cdot), \sigma(\cdot,\cdot)$, $\sigma^{-1}(\cdot,\cdot)$, and $g(\cdot,\cdot)$ are bounded and continuous, and the first two derivatives of $g(\cdot,\alpha)$ are bounded and continuous.

The Markov chain approximation. Theorem 8.1 below uses either the "central difference" or the "one-sided difference" schemes discussed in §5. The details will be worked out for the first case only for simplicity. Before stating the theorem, we will do some preliminary calculations that will be needed below. We have $a=\sigma^2$ and the transition probabilities

$$\begin{aligned}p^h(x\pm h|x,\alpha_0)&=[a(x,\alpha_0)\pm hb(x,\alpha_0)]/2a(x,\alpha_0)\\&=\frac{1}{2}\pm hb(x,\alpha_0)/(2a(x,\alpha_0)),\end{aligned}\tag{8.2}$$

where we assume that $\inf_{x,\alpha}[a(x,\alpha)-h|b(x,\alpha)|]>0$. This can always be guaranteed if we let h depend on x. Also, $\Delta t^h(x,\alpha)=h^2/a(x,\alpha)$. Then to get the v_i^h, which are used to define $Z^h(\cdot,\alpha_0)$, we need to evaluate the expression

$$\begin{aligned}\frac{p_\alpha^h(x\pm h|x,\alpha_0)}{p^h(x\pm h|x,\alpha_0)}&=\frac{[\pm hb_\alpha(x,\alpha_0)/a(x,\alpha_0)\mp hb(x,\alpha_0)a_\alpha(x,\alpha_0)/a^2(x,\alpha_0)]}{[1\pm hb(x,\alpha_0)/a(x,\alpha_0)]}\\&=\pm hb_\alpha(x,\alpha_0)/a(x,\alpha_0)\mp hb(x,\alpha_0)a_\alpha(x,\alpha_0)/a^2(x,\alpha_0)\\&\quad-h^2b(x,\alpha_0)b_\alpha(x,\alpha_0)/a^2(x,\alpha_0)\\&\quad+h^2b^2(x,\alpha_0)a_\alpha(x,\alpha_0)/a^3(x,\alpha_0)+O(h^3).\end{aligned}\tag{8.3}$$

The first and third terms on the right side of (8.3) also occurred in the example given at the beginning of §7. The second and fourth terms are new here. In the sums below, N_h is the number of steps required for the interpolated time to reach (modulo $O(h^2)$) the value T, when the interpolation intervals are $\{\Delta t^h(\xi_i^h,\alpha_0)\}$. Thus, for any parameter value,

$$\sum_{i=0}^{N_h-1}\Delta t^h(\xi_i^h,\alpha_0)=T+O(h^2).$$

N_h depends on the path, but not otherwise on α. In the previous sections, Δt^h did not depend on α. Using (8.3), we can write

$$\begin{aligned}\sum_{i=0}^{N_h-1}v_i^h&=\sum_{i=0}^{N_h-1}\left[\frac{b_\alpha(\xi_i^h,\alpha_0)}{a(\xi_i^h,\alpha_0)}-\frac{b(\xi_i^h,\alpha_0)a_\alpha(\xi_i^h,\alpha_0)}{a^2(\xi_i^h,\alpha_0)}\right]\delta\xi_i^h\\&\quad-\sum_{i=0}^{N_h-1}\left[\frac{b_\alpha(\xi_i^h,\alpha_0)b(\xi_i^h,\alpha_0)}{a(\xi_i^h,\alpha_0)}-\frac{b^2(\xi_i^h,\alpha_0)a_\alpha(\xi_i^h,\alpha_0)}{a^2(\xi_i^h,\alpha_0)}\right]\\&\quad\cdot\Delta t^h(\xi_i^h,\alpha_0)+O(h).\end{aligned}\tag{8.4}$$

There is a similar expression for the "one-sided difference" case, where we use the transition probabilities

$$p^h(x\pm h|x,\alpha)=\frac{a(x,\alpha)/2+hb^\pm(x,\alpha)}{a(x,\alpha)+h|b(x,\alpha)|}.$$

DEFINITIONS. To state the next theorem, we need the following definitions. Let $A(\alpha)$ denote the differential operator of the $x(\cdot)$ defined by (8.1). Define

$$\overline{Z}^h(T,\alpha_0) = \sum_{i=0}^{N_h-1} v_i^h,$$

$$\begin{aligned}\overline{Z}(T,\alpha_0) = &\int_0^T \left[\frac{b_\alpha(x(s),\alpha_0)}{a(x(s),\alpha_0)} - \frac{b(x(s),\alpha_0)a_\alpha(x(s),\alpha_0)}{a^2(x(s),\alpha_0)}\right] dx(s)\\ &- \int_0^T \left[\frac{b_\alpha(x(s),\alpha_0)b(x(s),\alpha_0)}{a(x(s),\alpha_0)} - \frac{b^2(x(s),\alpha_0)a_\alpha(x(s),\alpha_0)}{a^2(x(s),\alpha_0)}\right] ds,\end{aligned} \tag{8.5}$$

$$\delta T^h(\alpha_0) = \sum_{i=0}^{N_h-1} \frac{a_\alpha(\xi_i^h,\alpha_0)}{a(\xi_i^h,\alpha_0)}\Delta t^h(\xi_i^h,\alpha_0),$$

$$\delta T(\alpha_0) = \int_0^T \frac{a_\alpha(x(s),\alpha_0)}{a(x(s),\alpha_0)}\, ds,$$

$$\begin{aligned}\overline{Q}^h(\alpha_0) = &\overline{Z}^h(T,\alpha_0)C(\xi^h(\cdot),\alpha_0) + C_\alpha(\xi^h(\cdot),\alpha_0)\\ &- \sum_{i=0}^{N_h-1} k(\xi_i^h,\alpha_0)\frac{a_\alpha(\xi_i^h,\alpha_0)}{a(\xi_i^h,\alpha_0)}\Delta t^h(\xi_i^h,\alpha_0)\\ &+ k(\xi_{N_h}^h,\alpha_0)\delta T^h(\alpha_0) + A(\alpha_0)g(\xi_{N_h}^h,\alpha_0)\delta T^h(\alpha_0),\end{aligned} \tag{8.6}$$

$$\begin{aligned}\overline{Q}(\alpha_0) = &\overline{Z}(T,\alpha_0)C(x(\cdot),\alpha_0) + C_\alpha(x(\cdot),\alpha_0) - \int_0^T k(x(s),\alpha_0)\frac{a_\alpha(x(s),\alpha_0)}{a(x(s),\alpha_0)}\, ds\\ &+ k(x(T),\alpha_0)\delta T(\alpha_0) + A(\alpha_0)g(x(T),\alpha_0)\delta T(\alpha_0).\end{aligned} \tag{8.7}$$

THEOREM 8.1. *Assume* (A8.1) *and that* (8.1) *has a* (*weak sense*) *unique solution for each* $\alpha \in A_0$. *Use the central difference formula* (8.2). *Then* $V(\cdot)$ *is differentiable and*

$$(\xi^h(\cdot),\overline{Q}^h(\alpha_0)) \Rightarrow (x(\cdot),\overline{Q}(\alpha_0)),$$

and $\overline{Q}(\alpha_0)$ *is an unbiased estimator of* $V_\alpha(\alpha_0), \alpha_0 \in A_0$. (*There is a similar result for the "one-sided difference" case.*)

Remark on the proof. The procedure of §§5 and 6 can be followed, except for two details. The first is that we do not have an a priori formula for an estimator of $V_\alpha(\alpha_0)$ analogous to (3.3). Second, the interpolation time interval $\Delta t^h(x,\alpha) = h^2/a(x,\alpha)$ depends on α here. Hence the number of discrete timesteps that are needed for the interpolated time to reach the value T depends on α. To see the point more clearly, consider the special case where σ does not depend on x, and write $\Delta t^h(x,\alpha) = \Delta t^h(\alpha), a(x,\alpha) = a(\alpha)$. Then

$$\begin{aligned}\Delta t^h(\alpha_0 + \delta\alpha) - \Delta t^h(\alpha_0) &= \Delta t_\alpha^h(\alpha_0)\delta\alpha + o(\delta\alpha)\\ &= -\,\Delta t^h(\alpha_0)[a_\alpha(\alpha_0)]\delta\alpha + o(\delta\alpha).\end{aligned}$$

Note that $N_h = T/\Delta t^h(\alpha_0)$ here. Let $a_\alpha(\alpha_0) > 0$. The interpolated time reached in N_h discrete steps is short of reaching T by the amount $T[a_\alpha(\alpha_0)/a(\alpha_0)]\delta\alpha + o(\delta\alpha)$. The terms in (8.6) and (8.7) that have not appeared in $Q^h(\alpha_0)$ and $Q(\alpha_0)$ are due to the "compensation" necessitated by this "shortfall."

Proof. By the weak convergence argument in Theorem 7.1, $V^h(\alpha_0) \to V(\alpha_0)$ and $\xi^h(\cdot) \Rightarrow x(\cdot)$, satisfying (8.1). (The weak convergence assertions below are all provided by arguments similar to that of Theorem 7.1.) Let $N_h(\delta\alpha)$ denote the number of discrete timesteps required for the interpolated time for the chain $\{\xi_i^h\}$ under parameter $\alpha_0 + \delta\alpha$ to reach time T; i.e., $\sum_{i=0}^{N_h(\delta\alpha)-1} \Delta t^h(\xi_i^h, \alpha_0 + \delta\alpha) = T$ (mod $O(h^2)$). To simplify the proof, we let k and g not depend on α. The additional details that would be required are minor. We can write

$$\begin{aligned} &E^h_{\alpha_0+\delta\alpha} C(\xi^h(\cdot)) - E^h_{\alpha_0} C(\xi^h(\cdot)) \\ &= E^h_{\alpha_0+\delta\alpha}\left[\sum_{i=0}^{N_h(\delta\alpha)-1} k(\xi_i^h)\Delta t^h(\xi_i^h, \alpha_0+\delta\alpha) + g(\xi^h_{N_h(\delta\alpha)})\right] \\ &\quad - E^h_{\alpha_0}\left[\sum_{i=0}^{N_h-1} k(\xi_i^h)\Delta t^h(\xi_i^h, \alpha_0) + g(\xi^h_{N_h})\right]. \end{aligned} \tag{8.8}$$

Let P^h_α denote the measure induced by the first N_h steps of $\{\xi_i^h\}$ under parameter α. The $P^h_\alpha, \alpha \in A_0$, are mutually absolutely continuous, and the right side of (8.8) can be rewritten as

$$E^h_{\alpha_0} S^h_1(\delta\alpha) + E^h_{\alpha_0+\delta\alpha} S^h_2(\delta\alpha) + E^h_{\alpha_0+\delta\alpha} S^h_3(\delta\alpha) + E^h_{\alpha_0+\delta\alpha} S^h_4(\delta\alpha), \tag{8.9}$$

where

$$S^h_1(\delta\alpha) = \left[\sum_0^{N_h-1} k(\xi_i^h)\Delta t^h(\xi_i^h, \alpha_0) + g(\xi^h_{N_h})\right]\left(\frac{dP^h_{\alpha_0+\delta\alpha}}{dP^h_{\alpha_0}} - 1\right),$$

$$S^h_2(\delta\alpha) = \sum_{i=0}^{N_h-1} k(\xi_i^h)\left[\Delta t^h(\xi_i^h, \alpha_0+\delta\alpha) - \Delta t^h(\xi_i^h, \alpha_0)\right],$$

$$S^h_3(\delta\alpha) = \sum_{i=N_h}^{N_h(\delta\alpha)-1} k(\xi_i^h)\Delta t^h(\xi_i^h, \alpha_0+\delta\alpha),$$

$$S^h_4(\delta\alpha) = \left[g(\xi^h_{N_h(\delta\alpha)}) - g(\xi^h_{N_h})\right].$$

(Of course, we omitted the terms that would occur if k and g depended on α.) We use the summation conventions $\sum_a^b = -\sum_b^a$ if $a > b-1$, and $\sum_a^{a-1} = 0$.

Using the convergence $[dP^h_{\alpha_0+\delta\alpha}/dP^h_{\alpha_0} - 1]/\delta\alpha$ to $\sum_{i=0}^{N_h-1} v_i^h$ as $\delta\alpha \to 0$, we see that $S^h_1(\delta\alpha)/\delta\alpha$ converges to the first term on the right of (8.6), as $\delta\alpha \to 0$. Note that (A6.1) holds here. Then letting $h \to 0$ and using representation (8.4) and the results of Theorem 6.1, a weak convergence proof similar to that of Theorem 7.1 yields (parameter α_0)

$$(\xi^h(\cdot), \overline{Z}^h(T, \alpha_0)C(\xi^h(\cdot), \alpha_0)) \Rightarrow (x(\cdot), \overline{Z}(T, \alpha_0)C(x(\cdot), \alpha_0)).$$

Note that

$$\begin{aligned}(8.10)\qquad \Delta t^h(x,\alpha_0+\delta\alpha)-\Delta t^h(x,\alpha_0) &= \Delta t^h_\alpha(x,\alpha_0)\delta\alpha+o(\delta\alpha)\\ &= -h^2\frac{a_\alpha(x,\alpha_0)}{a^2(x,\alpha_0)}\delta\alpha+o(\delta\alpha)\\ &= -\Delta t^h(x,\alpha_0)\frac{a_\alpha(x,\alpha_0)}{a(x,\alpha_0)}\delta\alpha+o(\delta\alpha).\end{aligned}$$

Using (8.10), we get that $S_2^h(\delta\alpha)/\delta\alpha$ converges to the third term of (8.6) as $\delta\alpha\to 0$. Then another weak convergence argument yields that the third term of (8.6) converges weakly to the third term of (8.7) as $h\to 0$. We can write (modulo $O(h^2)$)

$$S_3^h(\delta\alpha)=\int_{T^h(\delta\alpha)}^{T} k(\xi^h(s))\,ds,$$

where $T^h(\delta\alpha)$ is the amount of interpolated time that passes in the first N_h steps for $\{\xi_n^h\}$, under parameter $\alpha_0+\delta\alpha$: i.e., where the interpolation intervals are $\{\Delta t^h(\xi_n^h,\alpha_0+\delta\alpha)\}$. It equals

$$\begin{aligned}T^h(\delta\alpha) &= \sum_{i=0}^{N_h-1}\Delta t^h(\xi_i^h,\alpha_0+\delta\alpha)\\ &= \sum_{i=0}^{N_h-1}[\Delta t^h(\xi_i^h,\alpha_0)+\Delta t^h_\alpha(\xi_i^h,\alpha_0)\delta\alpha+O(h^2)o(\delta\alpha)]\\ (8.11)\qquad &= T+\sum_{i=0}^{N_h-1}\Delta t^h_\alpha(\xi_i^h,\alpha_0)\delta\alpha+O(h^2)+o(\delta\alpha)\\ &= T-\sum_{i=0}^{N_h-1}\frac{a_\alpha(\xi_i^h,\alpha_0)}{a(\xi_i^h,\alpha_0)}\Delta t^h(\xi_i^h,\alpha_0)\delta\alpha+O(h^2)+o(\delta\alpha).\end{aligned}$$

Using (8.11), we have that $S_3^h(\delta\alpha)/\delta\alpha$ converges to the fourth term of $\overline{Q}^h(\alpha_0)$ as $\delta\alpha\to 0$. Also, that fourth term converges weakly to the fourth term of $\overline{Q}(\alpha_0)$ as $h\to 0$. A similar analysis shows that the $S_4^h(\delta\alpha)/\delta\alpha$ converges (modulo $O(h)$) to the last term of $\overline{Q}^h(\alpha_0)$ as $\delta\alpha\to 0$, and that this limit converges weakly to the last term of $\overline{Q}(\alpha_0)$ as $h\to 0$.

By the limits obtained above when $\delta\alpha\to 0$ (but not $h\to 0$), we have proved that

$$(8.12)\qquad V_\alpha^h(\alpha_0)=E_{\alpha_0}^h\overline{Q}^h(\alpha_0)+O(h).$$

In the proof, α_0 was fixed, but (8.12) holds for any $\alpha_0\in A_0$. This implies that if $[\alpha_0,\alpha_1]\in A_0$, then

$$(8.13)\qquad V^h(\alpha_1)-V^h(\alpha_0)=\int_{\alpha_0}^{\alpha_1}E_\beta^h\overline{Q}^h(\beta)\,d\beta+O(h).$$

We also have

$$E_\beta^h\overline{Q}^h(\beta)\to E_\beta\overline{Q}(\beta)$$

uniformly on $[\alpha_0, \alpha_1]$, and the right-hand side is continuous in β. The convergence follows from the weak convergence and the assertions above. The proof of the uniformity and continuity assertion is by contradiction. Suppose, for example, that the uniform convergence is false. Then there are $\beta, h_n \to 0$, $\beta_n \to \beta$, and $\delta_0 > 0$ such that

$$|E^{h_n}_{\beta_n}\overline{Q}^{h_n}(\beta_n) - E_\beta \overline{Q}(\beta)| \geq \delta_0.$$

However $\overline{Q}^{h_n}(\beta_n) \Rightarrow \overline{Q}(\beta)$, a contradiction. A similar argument yields the asserted β-continuity. The facts above and (8.13) imply that

$$V(\alpha_1) - V(\alpha_0) = \int_{\alpha_0}^{\alpha_1} E_\beta \overline{Q}(\beta)\, d\beta.$$

This yields the differentiability of $V(\cdot)$ as well as the fact that $\overline{Q}^h(\alpha_0)$ is an asymptotically unbiased estimator of $V_\alpha(\alpha_0)$. □

9. Reflected diffusion and heavy-traffic problems. In this section we discribe, rather loosely, some of the wide variety of models to which the results of §§3–8 can be extended.

Controlled reflected diffusions. The basic idea of §3 readily extends to "reflected problems," since the Girsanov transformation idea does. In the interest of saving space, we present a somewhat loose discussion. Consider a reflected diffusion modeled as a solution to the Skorokhod problem [13] in a set G satisfying the condition.

(A9.1) G is the closure of a bounded open set with a continuously differentiable boundary. Let $n(x)$ denote the outward normal to the boundary ∂G at x, and let $\beta(x)$ denote the reflection direction. Let $-\beta'(x)n(x) \geq \beta_0 > 0$, for all $x \in \partial G$.

The system is defined by (write $x(t) = x(t, \alpha)$, var= variation)

$$\begin{aligned} x(t) &= x + \int_0^t b(x(s), \alpha)\, ds + \int_0^t \sigma(x(s))\, dw\,(s) + Y(t), \\ (\text{var}\ \ Y)(t) \equiv |Y|(t) &= \int_0^t I_{\{x(s)\in\partial G\}} d|Y|(s), \\ Y(t) &= \int_0^t \beta(x(s)) d|Y|(s). \end{aligned} \tag{9.1}$$

We let b and σ be functions of the present state only for notational simplicity. The general case of §3 can also be handled. Let $x(\cdot)$, under parameter α, induce the measure P_α. Suppose that $\sigma^{-1}(\cdot)$ is bounded and continuous. (The degenerate case can also be treated, as in §3.) The P_α are mutually absolutely continuous and $dP_{\alpha_0+\delta\alpha}/dP_{\alpha_0}$ is that given by Theorem 3.1, where σ_2, w_2, and b_2 are replaced by σ, w, and b. The entire proof of Theorem 3.1 can be carried over. The only restriction is that the $\beta(\cdot)$ not depend on α (for then we lose the mutual absolute continuity). The cost C can be a function of both $x(\cdot)$ and $Y(\cdot)$. The discrete-time and Markov chain approximations of §§4–7 also work. Some appropriate Markov chains are discussed in [10].

Heavy traffic models. We now describe a model that has been widely used in the analysis of queueing systems under heavy traffic in that the appropriately scaled and normalized queue length processes converge weakly to it, as the traffic intensity converges to unity (i.e., as the relative idle time converges to zero) [14], [15]. There are K connected processors, and p_{ij} is the probability that an output from processor i goes to processor j. The matrix $P = \{p_{ij}\}$ has a spectral radius less than unity. Thus, all customers eventually leave the system. The parameter α can correspond to "marginal" service or arrival rates, or to other control parameters. Let $B^i < \infty$ be the scaled buffer size and restrain $x^i(t)$, the ith component of the state, to be in $[0, B^i]$.

The system is

$$x(t) = x + \int_0^t b(x(s), \alpha)\, ds + \int_0^t \sigma(x(s))\, dw\,(s) + (I - P')Y(t) - U(t), \tag{9.2}$$

where $Y^i(\cdot)$(respectively, $U^i(\cdot)$) are nondecreasing and can increase only at the t at which $x^i(t) = 0$ (respectively, $x^i(t) = B^i$). All the comments in the above section on controlled reflected diffusions apply here. Also, the cost C can be a function of $x(\cdot), Y(\cdot)$ and $U(\cdot)$.

These examples illustrate the power and range of applicability of the basic idea. Note also that the "mean square derivative" method described in §2 cannot be used for either of the cases of this section.

Generally, it is nearly impossible to obtain unbiased estimators of the derivatives with respect to α for the original physical queueing system whose heavy traffic limit is (9.2). For this reason, we might want to approximate the physical queue process by (9.2), and then compute or estimate the "sensitivities" for (9.2).

10. The small noise problem. For deterministic systems, the "deterministic form" of the mean square derivative method of §2 is often used, and it is of interest to know whether the estimates (3.3) or (3.5) converge to the "deterministic" derivative, as $\sigma \to 0$. This will be shown to be the case under appropriate conditions. This result is of more than theoretical interest, since it is often useful to add "small noise" and do the stochastic computation if less computation time is required to get a good estimate. More will be said about this below.

The following assumption will be used.

(A10.1) $b_i(\cdot)$, $k(\cdot)$, and $g(\cdot)$ are bounded and continuous and have bounded and continuous (x, α)-derivatives of first order and x-derivatives of second order. The σ_i are constant matrices, σ_2^{-1} exists, and the $w_i(\cdot)$ are mutually independent Wiener processes.

The b_i, k, and g could be "nonanticipative" functionals as in §3. Then, however, the "variational" or "linearization" equation (10.2) would be much more complicated. The results to be presented hold under any conditions that allow a linearization analogous to (10.2), (10.7).

The deterministic calculation. The deterministic problem is (parameter α)

$$\dot{x}_1 = b_1(x), \quad \dot{x}_2 = b_2(x, \alpha), \quad x \in R^r, \tag{10.1}$$

$$V(\alpha) = \int_0^T k(x(s), \alpha)\, ds + g(x(T), \alpha).$$

Set $x_{\alpha_0}(t) = \partial x(t)/\partial\alpha$, with $\alpha = \alpha_0$. Then (parameter α_0)

$$\dot{x}_{\alpha_0} = b_\alpha(x, \alpha_0) + b_x(x, \alpha_0)x_{\alpha_0}, \tag{10.2}$$

$$V_\alpha(\alpha_0) = \int_0^T [k_\alpha(x(t),\alpha_0) + k'_x(x(t),\alpha_0)x_{\alpha_0}(t)]\,dt$$
$$+ [g_\alpha(x(T),\alpha_0) + g'_x(x(T),\alpha_0)x_{\alpha_0}(T)].$$

Let $\Phi(s,t), s \le t$, be the fundamental solution of the ODE $\dot{y} = b_x(x(t),\alpha_0)y$. Until the end, we drop the $g(\cdot)$ for simplicity. We have

$$V_\alpha(\alpha_0) = \int_0^T dt \left\{ k_\alpha(x(t),\alpha_0) + k'_x(x(t),\alpha_0)[\int_0^t \Phi(u,t)b_\alpha(x(u),\alpha_0)du] \right\}. \tag{10.3}$$

The evaluation of (10.3) can be quite cumbersome, particularly for high-dimensional systems, and where α is a vector. We will show that $V_\alpha(\alpha_0)$ can be approximated by a slight variation of the stochastic formula (3.5) for small ϵ.

The small noise stochastic system will be (parameter α)

$$dx^\epsilon = \begin{pmatrix} b_1(x^\epsilon) \\ b_2(x^\epsilon,\alpha) \end{pmatrix} dt + \epsilon \begin{pmatrix} \sigma_1 & dw_1 \\ \sigma_2 & dw_2 \end{pmatrix}, \tag{10.4}$$
$$V^\epsilon(\alpha) = E^\epsilon_\alpha C(x^\epsilon(\cdot),\alpha),$$
$$C(x^\epsilon(\cdot),\alpha) = \int_0^T k(x^\epsilon(t),\alpha)\,dt + g(x^\epsilon(T),\alpha).$$

A stochastic calculation. From Theorem 3.1, we have

$$V^\epsilon_\alpha(\alpha_0) = \int_0^T k(x^\epsilon(t),\alpha_0)Z^\epsilon(t,\alpha_0)\,dt + \int_0^T k_\alpha(x^\epsilon(t),\alpha_0)\,dt, \tag{10.5}$$

where $Z^\epsilon(\cdot,\alpha_0)$ is just the $Z(\cdot,\alpha_0)$ of Theorem 3.1, with variance scale parameter ϵ used. The second right-hand integral in (10.5) converges in the mean to the corresponding term of (10.3) and will be omitted henceforth. We next set up the problem so that the appropriate estimator can be defined, and then the theorem will be stated.

Write $x^\epsilon(t) = x(t)+\tilde{x}^\epsilon(t)$, where $x(\cdot)$ solves (10.1). Then $\sup_{t\le T} E|\tilde{x}^\epsilon(t)| = O(\epsilon^2)$. Using the definition of $Z^\epsilon(\cdot,\alpha_0)$, the first term on the right side of (10.5) can be written as

$$\frac{1}{\epsilon}\int_0^T dtk(x(t)+\tilde{x}^\epsilon(t),\alpha_0)\int_0^t [\sigma_2^{-1}b_{2,\alpha}(x(s)+\tilde{x}^\epsilon(s),\alpha_0)]'dw_2(s). \tag{10.6}$$

Expand $k(x(t)+\tilde{x}^\epsilon(t),\alpha_0) = k(x(t),\alpha_0)+k'_x(x(t),\alpha_0)\tilde{x}^\epsilon(t)+O(|\tilde{x}^\epsilon(t)|^2)$ and write (10.6) as

$$T^\epsilon_1 + T^\epsilon_2 + O(\epsilon^2),$$

where $E|O(\epsilon^2)| = O(\epsilon^2)$ and

$$T^\epsilon_1 = \frac{1}{\epsilon}\int_0^T dtk(x(t),\alpha_0)\int_0^t [\sigma_2^{-1}b_{2,\alpha}(x^\epsilon(s),\alpha_0)]'dw_2(s) = \int_0^T dt\; k(x(t),\alpha_0)Z^\epsilon(t,\alpha_0),$$

$$T^\epsilon_2 = \frac{1}{\epsilon}\int_0^T dtk'_x(x(t),\alpha_0)\tilde{x}^\epsilon(t)\int_0^t [\sigma_2^{-1}b_{2,\alpha}(x(s)+\tilde{x}^\epsilon(s),\alpha_0)]'dw_2(s).$$

Note that $ET_1^\epsilon = 0$, since $x(\cdot)$ is deterministic. Thus, we need not include T_1^ϵ in the estimator (also, its variance is $O(1/\epsilon^2)$, in general). We can write $\tilde{x}^\epsilon(\cdot)$ as

$$d\tilde{x}^\epsilon = b_x(x(t), \alpha_0) \cdot \tilde{x}^\epsilon dt + \epsilon\sigma dw + O(|\tilde{x}^\epsilon|^2)\, dt, \qquad \tilde{x}^\epsilon(0) = 0, \tag{10.7a}$$

or, equivalently,

$$\tilde{x}^\epsilon(t) = \epsilon \int_0^t \Phi(u,t)\sigma\, dw\,(u) + \delta_\epsilon(t), \tag{10.7b}$$

where $\sup_{t\leq T} E|\delta_\epsilon(t)| = O(\epsilon^2)$.

Next, expand the $b_{2,\alpha}$ in T_2^ϵ and substitute (10.7b) into it to yield

$$\begin{aligned} T_2^\epsilon = \int_0^T dt k_x'(x(t),\alpha_0) \int_0^t \Phi(u,t)\sigma\, dw\,(u) \\ \cdot \int_0^t [\sigma_2^{-1} b_{2,\alpha}(x(s),\alpha_0)]'\, dw_2\,(s) + O(\epsilon), \end{aligned} \tag{10.8}$$

where $E|O(\epsilon)| = O(\epsilon)$. The mean value of the primary term in (10.8) is

$$\int_0^T dt k_x'(x(t),\alpha_0) \int_0^t \Phi(u,t) b_\alpha(x(u),\alpha_0)\, du,$$

which is just (10.3) (with the k_α term dropped there).

Reintroducing the k_α term, and doing a similar expansion for the g component of the cost, the above discussion can be summarized by the following theorem.

THEOREM 10.1. *Assume* (A10.1). *Then* $V_\alpha^\epsilon(\alpha_0) \to V_\alpha(\alpha_0)$ *as* $\epsilon \to 0$. *Define*

$$\begin{aligned} \tilde{Q}^\epsilon(\alpha_0) = \left[\int_0^T k(x^\epsilon(t),\alpha_0) Z^\epsilon(t,\alpha_0)\, dt - \int_0^T k(x(t),\alpha_0) Z^\epsilon(t,\alpha_0) dt\right] \\ + \int_0^T k_\alpha(x(t),\alpha_0)\, dt + g_\alpha(x(T),\alpha_0) \\ + \left[g(x^\epsilon(T),\alpha_0) - g(x(T),\alpha_0)\right] Z^\epsilon(T,\alpha_0). \end{aligned} \tag{10.9}$$

Then $E\tilde{Q}^\epsilon(\alpha_0) \to V_\alpha(\alpha_0)$ *as* $\epsilon \to 0$. *The variance of each term is bounded uniformly in* ϵ.

The primary computational advantage in using (10.9) as a "surrogate" estimate of $V_\alpha(\alpha_0)$ is that we need not calculate $\Phi(\cdot,\cdot)$, an onerous task, nor evaluate the integrals in (10.3), separately for each component if α is a vector. The integrals in (10.9) are usually much easier to calculate than those in (10.3), and they require only a single simulation with no noise and one with small noise. If the variance of (10.9) is not too large, then it might be computationally advantageous to simulate the stochastic system several times and average the results, rather than work with (10.3). The preferable method depends on the case, however; see the next section.

11. Some numerical comparisons. Numerical comparisons of the methods of §§4 and 5 with the methods of §2 have been made on a number of problems, linear and nonlinear, and for both high and low dimension. Definitive comparisons are difficult, since each method allows different possibilities for variance reduction and realization.

Thus the results below are taken as only tentative. The runs for the finite-difference method were taken using common random numbers for all the runs needed to get a single finite-difference estimate of the gradient or derivative. This does provide some reduction in the variance of the estimators. For the methods of §§4 and 5, two approaches were used. The first used "independent" random variables for each sample run. The second (called AV for antithetic variable) took the sample runs in pairs, the random variables being "independent" from pair to pair, with those used in one member of the pair being the negative of those used in the other member of the pair. This did yield a substantial reduction in the variance of the estimators, as can be seen from the tabulated data below. For all cases 5,000 samples were taken.

Linear systems. If the system is linear and the cost function smooth, then the classical methods of §2 seem to have the advantage over the methods of §§4 and 5, at least for the systems simulated. The reasons are not clear (although, due to the linearity, we were able to simplify some of the calculations for the former methods), but all the methods should be treated as serious competitors until data indicates otherwise. The former methods used more computation time, as expected, but the variance of the estimates was sufficiently smaller, so as to give those methods the advantage (sometimes a clear advantage, other times only a slight advantage). The values of the finite-difference intervals were not too important (within reason), for the particular simulation method and problem class used. Also, the finite difference and the "mean square derivative" methods performed about the same, in terms of both CPU time and the quality of the estimates. The addition of noise to more than one state or using more complex dynamics shifts the preference.

Nonlinear systems. The numerical comparisons were more informative for nonlinear systems with "nonsmooth" cost functions. The data in Tables 1–5 was taken for the system

$$\ddot{x} - \dot{x}(1 - \alpha x^2) - x = .5 \quad \text{(white noise)} \tag{11.1}$$

with $x(0) = 1$, $\dot{x}(0) = 1$, $\alpha = 0.5$, and the cost function

$$P\{\sup_{2\geq t\geq 0} |x(t)| \geq 3.8\}. \tag{11.2}$$

Here the parameter is scalar valued and, for the finite-difference method, the difference interval is denoted by $\delta\alpha$, and the system is simulated by a discrete-time approximation of the type used in §4 with a time-difference interval of .01. Simulations indicate that the true value of the derivative is between -4.8 and -5.1 with a very high probability. We see from Table 1 that the value chosen for $\delta\alpha$ is important, and a substantial bias results if the value is too big. If the time interval is increased to .05, then the CPU time for the finite-difference case would be reduced to about 4 with essentially the same biases and variances. Comparing Tables 1 and 4, the advantage of the method of §4 over the finite difference method is apparent, particularly if bias is a concern. Note also the advantage provided by the antithetic variable (AV) method, by comparing Tables 2 and 3. We would expect that the differences would be more pronounced for higher-dimensional nonlinear problems. Similarly, comparing Tables 1 and 5 (with the cited adjustment for the time-difference interval to .05) shows the advantage of the method of §5 to the finite-difference method. The mean square derivative method is not applicable here, since the cost function $C(x(\cdot))$ used in (11.2) is not differentiable. Tables 3(a) and 5(a) use both antithetic variables and the centering (3.5′). The given CPU times are for 5,000 estimates of the derivatives.

These results should be used with caution, since all methods can undoubtably be improved. Much depends on the difficulty of approximating the differential equation, and we can find meaningful problems where any given method performs best.

In an application to stochastic approximation, we might be less concerned with the biases in the first few iterates than in the later iterates, and the particular method that is used might vary, according to need.

TABLE 1.
The finite difference method.

	$\delta\alpha = .05$		$\delta\alpha = .01$	
	mean	variance	mean	variance
derivative	-2.704	46.8	-4.82	458.8
cost	.148	.126	.148	.126
CPU time	20.68			

TABLE 2.
The method of §4.

	$\Delta = .01$		$\Delta = .05$	
	mean	variance	mean	variance
derivative	-5.03	207	-5.05	210
cost	.148	.126	.151	.128
CPU time	slightly more than in Table 3			

TABLE 3.
The method of §4 (AV).

	$\Delta = .01$		$\Delta = .05$	
	mean	variance	mean	variance
derivative	-4.87	91	-4.96	93
cost	.145	.051	.149	.05
CPU time	12.73		2.54	

TABLE 3a.
The method of §4 (AV *and centering*).

	$\Delta = 0.01$		$\Delta = 0.05$	
	mean	variance	mean	variance
derivative	−4.83	81.58	−4.82	78.32
cost	0.143	0.051	0.149	0.05
CPU time	12.47		2.48	

TABLE 4.
The method of §5.

	$h = .1$		$h = .05$		$h = .02$	
	mean	variance	mean	variance	mean	variance
derivative	−3.7	76.4	−3.72	104	−4.5	152
cost	.227	.175	.165	.138	.163	.055
CPU time	4.21		13.86		77.5	

TABLE 5.
The method of §5 (AV).

	$h = .1$		$h = .05$	
	mean	variance	mean	variance
derivative	−3.75	42	−3.77	46
cost	.220	.08	.163	.055
CPU time	4		13.19	

TABLE 5a.
The method of §5 (AV *and centering*).

	$h = .1$		$h = .05$	
	mean	variance	mean	variance
derivative	−3.31	25.49	−3.78	39.42
cost	.183	0.06	0.163	0.055
CPU time	4.06		13.19	

REFERENCES

[1] F. Campillo, *Optimal ergodic control of nonlinear stochastic systems*, INRIA Report, 1989; Effective Stochastic Analysis, P. Krée and W. Wedig, eds., to appear.

[2] F. Campillo, F. LeGland, and E. Pardoux, *Approximation d'un problèm de contrôl ergodique dégénéré*, in New Trends in Nonlinear Control Theory, J. Descusse, M. Fliess, A. Isidori, and D. Leborgne, eds., Lecture Notes in Control and Inform. Sci., 122, Springer-Verlag, pp. 379–395, 1989.

[3] M. I. Reiman and A. Weiss, *Sensitivity analysis for simulations via likelihood ratios*, Oper. Res., 37 (1989), pp. 830–844.

[4] Y. C. Ho and C. Cassandras, *A new approach to the analysis of discrete event dynamical systems*, Automatica, 19 (1983), pp. 149–167.

[5] P. Heidelberger, Xi-Ren Cao, M. A. Zazanis, and R. Suri, *Convergence properties of infinitesimal perturbation analysis estimates*, Management Sci., 34 (1988), pp. 1281–1302.

[6] N. Ikeda and S. Watanabe, *Stochastic Differential Equations and Diffusion Processes*, North-Holland, Amsterdam, 1981.

[7] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[8] T. G. Kurtz, *Approximation of Population Processes*, CBMS-NSF Regional Conf. Ser. in Appl. Math., Vol. 36, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1981.

[9] H. J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.

[10] ———, *Numerical methods for stochastic control problems in continuous time*, SIAM J. Control Optim., 28 (1990), pp. 999–1048, 1990.

[11] H. J. Kushner and G. B. DiMasi, *Approximations for functionals and optimal control problems on jump-diffusion processes*, J. Math. Anal. Appl., 63 (1978), pp. 772–800.

[12] H. Kunita and S. Watanabe, *On square integrable martingales*, Nagoya Math. J., pp. 209-245, 1967.

[13] P. L. Lions and A. S. Sznitman, *Stochastic differential equations with reflecting boundary conditions*, Comm. Pure Appl. Math., 37 (1984), pp. 511–553.

[14] H. J. Kushner and K. M. Ramachandran, *Optimal and approximately optimal control policies for queues in heavy traffic*, SIAM J. Control Optim., 27 (1989), pp. 1295–1318.

[15] M. I. Reiman, *Open queueing networks in heavy traffic*, Math Oper. Res., 9 (1984), pp. 441–458.

[16] E. Pardoux and D. Talay, *Discretization and simulation of stochastic differential equations*, Acta Appl. Math., 3 (1985), pp. 23–47.

[17] I. I. Gikhman and A. V. Skorohod, *Introduction to the Theory of Random Processes*, W. B. Saunders, Philadelphia, PA, 1969.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 6, pp. 1251-1287, November 1991

001

BALANCED PARAMETRIZATION OF CLASSES OF LINEAR SYSTEMS*

RAIMUND OBER†

Abstract. Canonical forms and parametrizations are presented for several sets of minimal systems of given dimension: asymptotically stable systems, allpass systems, bounded real systems, positive real systems, minimum-phase systems, and the class of all minimal systems. The approach is based on balancing techniques for these classes of systems. Applications of these results to Hankel operators and model reduction are discussed.

Key words. canonical form, parametrization, balanced realization, model reduction

AMS(MOS) subject classifications. 93B10, 93B20

1. Introduction and notation. Canonical forms for linear systems are of importance since they provide a unique state-space representation of linear systems. They therefore play a major role in system identification where a unique parametrization of the systems in the model set is essential. From a more theoretical point of view, canonical forms permit the study of topological and geometric properties of sets of linear systems [13], [15], [24]. For a survey of results and applications of canonical forms, see [15].

A definition of a canonical form is as follows.

DEFINITION 1.1. Let M be a set of minimal state-space systems. Then a map

$$\Gamma : M \mapsto M$$

is called a *canonical form* if

(1) $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) := \Gamma((A, B, C, D))$ is *system equivalent* to (A, B, C, D), i.e., there exists $T \in \Re^{n \times n}$, invertible such that

$$\tilde{A} = TAT^{-1}, \quad \tilde{B} = TB, \quad \tilde{C} = CT^{-1}, \quad \tilde{D} = D.$$

(2) If (A_1, B_1, C_1, D_1) and (A_2, B_2, C_2, D_2) are system equivalent, then

$$\Gamma((A_1, B_1, C_1, D_1)) = \Gamma((A_2, B_2, C_2, D_2)).$$

Various types of canonical forms for linear systems have been introduced and studied (see, e.g., [30], [12], [15]). Most of these canonical forms for multivariable systems are generalizations of the observer or controller canonical form for single-input single-output systems. The canonical forms presented here are based on balanced representations of linear systems. Balanced realizations for asymptotically stable systems were introduced in [19]. Other balancing techniques were then investigated in [7] for stochastic realizations, in [16] for the class of all minimal systems, and in [28] for bounded real systems. The motivation of those authors was to obtain a simple method for the approximation of a system by a lower-dimensional system in the same class.

It is shown in this paper that the balancing technique leads to canonical forms with desirable properties. In § 2 we consider balanced realizations as they were defined in [19] for asymptotically stable systems. The canonical form established for this class of systems is a modification of a similar canonical form presented in [22]. This canonical form has a structure that shows that allpass systems, which are considered in § 3, are

* Received by the editors November 16, 1988, accepted for publication (in revised form) September 10, 1990.

† Department of Engineering, Cambridge University, Trumpington Street, Cambridge CB1 1PZ, United Kingdom. Present address, Center for Engineering Mathematics, The University of Texas at Dallas, Richardson, Texas 75083.

in some sense building blocks of this canonical form. These results are the basis for the derivation of canonical forms of other classes of systems as presented in later sections of the paper. In [27] the canonical form in [22] was used to derive a canonical form for the class of minimal systems in terms of Riccati-balanced coordinates as introduced by [16]. This canonical form is rewritten in § 4 using the canonical form of § 2. Relating the class of bounded real systems to a subclass of multivariable asymptotically stable systems, it is possible to derive a canonical form for bounded real systems as presented in § 5. Via a Moebius-type transformation, the class of bounded real systems is mapped to the class of positive real systems. Since positive real systems are closely related to minimum-phase systems through spectral factorization, the canonical form derived for positive real systems in § 6 can be used in § 7 to derive a canonical form for minimum-phase systems. Section 8 deals with the relationship of the previous results to discrete-time systems. The question of model reduction of systems given in the canonical forms is considered in § 9. Examining the canonical forms, it appears that they have many common structural properties, which are discussed in § 10.

The objective of the paper is to show that for many classes of systems it is possible to derive canonical forms using the idea of balancing. The general principle of balancing is to associate with a particular class of systems a set of Riccati equations that are intrinsically related with the properties of the particular class of systems. The class of asymptotically stable systems is, for example, associated with a set of Lyapunov equations, whereas the set of positive real systems is associated with the positive real Riccati equations. It is then possible to define what a balanced realization means for such a class of systems. A system is called balanced if the solutions to the two associated equations are identical and diagonal. Having defined the notion of a balanced realization, it can be seen that such a realization of a particular system is not unique. One of the aims of this paper is to show that by imposing further constraints on the realization, it is indeed possible to obtain a unique realization, i.e., a canonical form.

The usefulness of canonical forms very much depends on their properties. One of the standard canonical forms, the controller canonical form, is of particular significance since the parameters of the canonical form have an immediate interpretation as the coefficients of the transfer function. There are, however, drawbacks of the controller canonical form especially concerning the resulting parametrization of linear systems. The set of parameters in the controller canonical form that lead to a minimal system is very complicated. This makes it difficult to use this canonical form in cases where it is necessary to have a geometrically well-behaved parameter space, e.g., in some optimization tasks. One of the main advantages of the canonical forms derived here is that it can be shown that the parameter spaces associated with a canonical form have—especially in the case of single-input single-output systems—desirable geometric properties. This is at the expense of having to partition the set of parametrized systems into suitable subsets. This means that even in the case of single-input single-output systems structural parameters must be introduced.

There would be several ways to derive the results presented here. One of them would be to treat each of the classes of systems separately and construct the canonical form from first principles. This approach would be very tedious, especially because of the complexity of the canonical forms in the case of multivariable systems. Instead, we are going to relate the various classes of systems to one another. This allows us to carry the canonical form over from one class of systems to another without having to repeat the basic construction.

Analyzing these canonical forms, it becomes apparent that all share certain structural properties. It is interesting to see that such widely differing classes of transfer functions, such as, for example, minimum-phase systems and bounded real systems, admit a parametrization that has very similar properties. Having a common structure has the advantage of allowing us to deal with the various classes of systems in a unified way. It was possible to exploit this common structure in the study of the connectivity properties of the various classes of systems [26].

For the presentation of our results we will use the common structure of the canonical forms. We will introduce the following notation, which allows us to simplify the statement of the canonical forms especially in the case of multivariable systems. Most of these definitions are, however, not important for the case of single-input single-output systems.

- A matrix $B=(b_{i,j})_{\substack{1\leqq i<k\\ 1\leqq j\leqq l}}$ is called *positive upper triangular* if there exist indices

$$1\leqq t_1<\cdots<t_j<\cdots<t_k\leqq l$$

such that

$$\begin{aligned} &b_{i,t_i}>0 && \text{for } 1\leqq i\leqq k,\\ &b_{i,j}=0 && \text{for } 1\leqq j<t_i \text{ and } 1\leqq i\leqq k,\\ &b_{i,j}\in\mathfrak{R} && \text{otherwise,}\end{aligned}$$

i.e.,

$$B=\begin{pmatrix} 0 & \cdots & 0 & b_{1,t_1} & b_{1,t_1+1} & \cdots & \cdot & \cdot & \cdot & \cdots & & & \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & b_{2,t_2} & b_{2,t_2+1} & \cdots & & & \\ \vdots & & \vdots & \vdots & \vdots & & \vdots & & & & & & \\ 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 & b_{k,t_k} & b_{k,t_k+1}\cdots \end{pmatrix}.$$

- A matrix A is said to be in *r-balanced form*, $1\leqq r\leqq n$, if for

$$A=\begin{pmatrix} A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix},\qquad A_{11}\in\mathfrak{R}^{r\times r},$$

we have
(1) A_{11} is skew symmetric.
(2) A_{12} and A_{22} are given by the set of indices

$$\begin{aligned} &1=h_1<\cdots<h_i<h_{i+1}<\cdots<h_q\leqq n-r,\\ &1\leqq g_q<\cdots<g_{i+1}<g_i<\cdots<g_1\leqq r\end{aligned}$$

in the following way:

(a) for $A_{12}=(a_{st})_{\substack{1\leqq s\leqq r\\ 1\leqq t\leqq n-r}}$ we have

$$\begin{aligned} &a_{g_i,h_i}>0 && \text{for } 1\leqq i\leqq q,\\ &a_{g_i,t}=0 && \text{for } t>h_i \quad \text{where } 1\leqq i\leqq q,\\ &a_{s,t}=0 && \text{for } t\geqq h_i \text{ and } s>g_i \quad \text{where } 1\leqq i\leqq q,\end{aligned}$$

i.e.,

$$A_{12}=\begin{pmatrix} \vdots & \vdots & & \vdots & \vdots & \vdots & \\ \cdot & \cdot & \cdots & a_{g_2-1,h_2-1} & a_{g_2-1,h_2} & a_{g_2-1,h_2+1} & \cdots \\ \cdot & \cdot & \cdots & a_{g_2,h_2-1} & a_{g_2,h_2} & 0 & \cdots \\ \cdot & \cdot & \cdots & a_{g_2+1,h_2-1} & 0 & 0 & \cdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \\ a_{g_1-1,h_1} & a_{g_1-1,h_1+1} & \cdots & a_{g_1-1,h_2-1} & 0 & 0 & \cdots \\ a_{g_1,h_1} & 0 & \cdots & 0 & 0 & 0 & \cdots \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \\ \vdots & \vdots & & \vdots & \vdots & \vdots & \\ 0 & 0 & \cdots & 0 & 0 & 0 & \cdots \end{pmatrix}.$$

(b) A_{22} is given by

$$A_{22}=\begin{pmatrix} 0 & \alpha_2 & & & & \\ -\alpha_2 & 0 & \alpha_3 & & & \\ & -\alpha_3 & 0 & \cdot & & 0 \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \cdot & 0 & \alpha_{n-r} \\ & & & & -\alpha_{n-r} & 0 \end{pmatrix},$$

where for $2\leqq i\leqq n-r$

$$\alpha_i\begin{cases} =0 & \text{if } i=h_s \text{ for some } 1\leqq s\leqq q, \\ >0 & \text{otherwise.} \end{cases}$$

(3) $A_{21}=-A_{12}^T$.

- Let $A=(a_{ij})_{\substack{1\leqq i\leqq n\\ 1\leqq j\leqq n}}$ then we denote by

(1) $[A]_l=(\bar{a}_{ij})_{\substack{1\leqq i\leqq n\\ 1\leqq j\leqq n}}$ the *lower triangular part* of A, i.e.,

$$\bar{a}_{ij}=\begin{cases} 0 & \text{for } j\geqq i, \\ a_{ij} & \text{for } j<i. \end{cases}$$

(2) $[A]_d=(\tilde{a}_{ij})_{\substack{1\leqq i\leqq n\\ 1\leqq j\leqq n}}$ the *diagonal part* of A, i.e.,

$$\tilde{a}_{ij}=\begin{cases} 0 & \text{for } j\neq i, \\ a_{ij} & \text{for } j=i. \end{cases}$$

- Let $(A, B, C, D)\in\mathfrak{R}^{n\times n}\times\mathfrak{R}^{n\times m}\times\mathfrak{R}^{p\times n}\times\mathfrak{R}^{p\times m}$ and $n_1,\cdots,n_j,\cdots,n_k$, $n_j\in\mathcal{N}$, $\sum_{j=1}^k n_j=n$. Then (A, B, C, D) is said to be *partitioned according to*

$n_1, \cdots, n_j, \cdots, n_k$ if

$$A = (A_{ij})_{1 \leq i,j \leq k}, \qquad A_{ij} \in \Re^{n_i \times n_j},$$

$$B = \begin{pmatrix} B_1 \\ \vdots \\ B_j \\ \vdots \\ B_k \end{pmatrix}, \qquad B_j \in \Re^{n_j \times m},$$

$$C = (C_1, \cdots, C_j, \cdots, C_k), \qquad C_j \in \Re^{p \times n_j}.$$

The following notation and abbreviations will be used throughout the paper.

- Classes of transfer functions:
  - —$TL_n^{p,m} = \{p \times m$ transfer functions of McMillan degree $n\}$.
  - —$TC_n^{p,m} = \{G(s) \in TL_n^{p,m} | G(s)$ has all its poles in the open left halfplane$\}$.
  - —$TA_n^m = \{G(s) \in TC_n^{m,m} | G(s)G(-s)^T = \sigma^2 I$, for some $\sigma > 0$, for all $s \in \mathscr{C}\}$.
  - —$TP_n^m = \{G(s) \in TC_n^{m,m} | G(\infty) + G(-\infty)^T > 0$, $G(iw) + G(-iw)^T > 0$ for all $w \in \Re\}$.
  - —$TB_n^{p,m} = \{G(s) \in TC_n^{p,m} | G(-\infty)^T G(\infty) < I$, $I - G(-iw)^T G(iw) > 0$ for all $w \in \Re\}$.
  - —$TM_n^m = \{G(s) \in TC_n^{m,m} | G(s)^{-1} \in TC_n^{m,m}\}$.
  - —The corresponding sets of discrete-time systems are defined in § 8.
- Classes of state-space systems:
  - —The sets of minimal state-space realizations of the transfer functions in $TL_n^{p,m}$, $TC_n^{p,m}$, TA_n^m, TP_n^m, $TB_n^{p,m}$, and TM_n^m are denoted by $L_n^{p,m}$, $C_n^{p,m}$, A_n^m, P_n^m, $B_n^{p,m}$, and M_n^m.
  - —The corresponding sets of discrete time systems are defined in § 8.
- Symbols:
  - —diag $(A_1, \cdots, A_k)$ is a block diagonal matrix with $A_1, \cdots, A_k$ as its block diagonal entries.
  - —$\hat{I}_n :=$ diag $(+1, -1, +1, -1, \cdots, (-1)^{n+1}) \in \Re^{n \times n}$.
  - —$\mathscr{N}$ denotes the set of natural numbers, $\mathscr{C}$ denotes the set of complex numbers, and $\mathscr{R}$ the set of real numbers.

2. Asymptotically stable systems. In this section we will review some of the results on the balanced parametrization of asymptotically stable systems as given in [22]. The particular canonical form presented here for multivariable systems is, however, a simplified version of the one introduced in [22]. First, we will define a balanced realization of a minimal and asymptotically stable continuous-time system as introduced by Moore [19].

DEFINITION 2.1. The set of minimal and asymptotically stable systems $(A, B, C, D) \in \Re^{n \times n} \times \Re^{n \times m} \times \Re^{p \times n} \times \Re^{p \times m}$ is denoted by $C_n^{p,m}$. A system $(A, B, C, D) \in C_n^{p,m}$ is called *balanced* if there exists a diagonal matrix $\Sigma =$ diag $(\sigma_1, \cdots, \sigma_j, \cdots, \sigma_n)$ such that

$$(1) \qquad A\Sigma + \Sigma A^T = -BB^T, \qquad A^T\Sigma + \Sigma A = -C^T C.$$

The matrix Σ is called the gramian of the system (A, B, C, D) and its diagonal entries are said to be the *singular values* of the system.

Moore [19] has shown that each system in $C_n^{p,m}$ has an equivalent system that is balanced. Such a realization, however, is not unique. If $\Sigma =$ diag $(\sigma_1 I_{n(1)}, \cdots, \sigma_j I_{n(j)}, \cdots, \sigma_k I_{n(k)})$, $\sigma_1 > \cdots > \sigma_j > \cdots > \sigma_k > 0$, is the gramian of a balanced system (A, B, C, D), then all equivalent balanced systems with singular values ordered according to multiplicities are given by (QAQ^T, QB, CQ^T, D), where $Q = \operatorname{diag}(Q_1, \cdots, Q_j, \cdots, Q_k)$, $Q_j \in \Re^{n_j \times n_j}$, $Q_j^T Q_j = I_{n_j}$, for $1 \leqq j \leqq k$ (see, e.g., [19], [10]). Thus if Σ has distinct diagonal entries, (A, B, C, D) is unique up to a state-space transformation by a sign matrix, i.e., a diagonal matrix whose diagonal entries are ± 1. Hence for this case a canonical form can be obtained by constraining the first nonzero entry of each row of the B-matrix to be positive [17], [20], [22]. The following theorem gives a canonical form for all systems in $C_n^{p,m}$ in terms of balanced realizations. It thereby shows how to impose further constraints to obtain a unique balanced realization in the general case.

THEOREM 2.1. *The following two statements are equivalent*:

(1) $G(s) \in TC_n^{p,m}$.

(2) $G(s)$ *has a realization* $(A, B, C, D) \in \Re^{n \times n} \times \Re^{n \times m} \times \Re^{p \times n} \times \Re^{p \times m}$ *given by the parameters*:

$$\begin{array}{llll}
\sigma_1 > \cdots > \sigma_j > \cdots > \sigma_k > 0 & & & \\
n_1, \cdots, & n_j, \cdots, & n_k, & n_j \in \mathcal{N}, \sum_{j=1}^k n_j = n; \\
r_1, \cdots, & r_j, \cdots, & r_k, & r_j \in \mathcal{N}, \quad 1 \leqq r_j \leqq \min(n_j, m, p); \\
U_1, \cdots, & U_j, \cdots, & U_k, & U_j \in \Re^{p \times r_j}, \quad U_j^T U_j = I_{r_j}; \\
\tilde{B}_1, \cdots, & \tilde{B}_j, \cdots, & \tilde{B}_k, & \tilde{B}_j \in \Re^{r_j \times m} \textit{ positive upper triangular}; \\
\tilde{A}_1, \cdots, & \tilde{A}_j, \cdots, & \tilde{A}_k, & \tilde{A}_j \in \Re^{n_j \times n_j} \textit{ in } r_j\textit{-balanced form}; \\
D, & & & D \in \Re^{p \times m};
\end{array}$$

in the following way.

If (A, B, C, D) *is partitioned according to* $n_1, \cdots, n_j, \cdots, n_k$, *then*

(i) $B_j = \begin{pmatrix} \tilde{B}_j \\ 0 \end{pmatrix}, \qquad 1 \leqq j \leqq k,$

(ii) $C_j = (U_j \Delta_j, 0)$ *where* $\Delta_j = (\tilde{B}_j \tilde{B}_j^T)^{1/2}, \qquad 1 \leqq j \leqq k,$

(iii) $A_{jj} = \tilde{A}_j - \dfrac{1}{\sigma_j} [\operatorname{diag}(\Delta_j^2, 0)]_l - \dfrac{1}{2\sigma_j} [\operatorname{diag}(\Delta_j^2, 0)]_d, \qquad 1 \leqq j \leqq k,$

(iv) $A_{ij} = \dfrac{1}{\sigma_i^2 - \sigma_j^2} \operatorname{diag}((\sigma_j \tilde{B}_i \tilde{B}_j^T - \sigma_i \Delta_i U_i^T U_j \Delta_j), 0), \qquad 1 \leqq i, j \leqq k, \quad i \neq j,$

(v) $D \in \Re^{p \times m}$.

Moreover, (A, B, C, D) *as defined in* (2) *is balanced with gramian*

$$\Sigma = \operatorname{diag}(\sigma_1 I_{n_1}, \cdots, \sigma_j I_{n_j}, \cdots, \sigma_k I_{n_k}).$$

The map Γ, *which assigns to each system in* $C_n^{p,m}$ *the realization given in* (2), *is a canonical form.*

Proof. The derivation of the results is similar to the derivation of the canonical form for systems in $C_n^{p,m}$ as given in [22].

(1) $\Rightarrow$ (2) Let $(A, B, C, D) \in C_n^{p,m}$ be a balanced system with gramian $\Sigma = \sigma I_n$. Then it follows from the Lyapunov equations (1) that

$$A + A^T = -\frac{1}{\sigma} BB^T = -\frac{1}{\sigma} C^T C.$$

Since the realization (A, B, C, D) is unique up to a state-space transformation with a unitary matrix, we can assume without loss of generality that

$$B=\begin{pmatrix}\tilde{B}\\ 0\end{pmatrix},$$

where $\tilde{B}\in\Re^{r\times m}$, $r=\operatorname{rank}(B)$, is in positive upper triangular form. Note that this is a unique representation of B.

Since $BB^T=C^TC$ there exists a unique $U\in\Re^{p\times r}$, $U^TU=I_r$, such that

$$C=(U(\tilde{B}\tilde{B}^T)^{1/2},0).$$

If we write

$$A=\begin{pmatrix}A_{11} & A_{12}\\ A_{21} & A_{22}\end{pmatrix},\qquad A_{11}\in\Re^{r\times r},$$

we have

$$A_{11}+A_{11}^T=-\frac{1}{\sigma}\tilde{B}\tilde{B}^T.$$

Thus the diagonal elements of A_{11} are given by

$$[A_{11}]_d=-\frac{1}{2\sigma}[\tilde{B}\tilde{B}^T]_d.$$

Since

$$A_{11}=-\frac{1}{\sigma}\tilde{B}\tilde{B}^T-A_{11}^T,$$

A_{11} is completely parametrized by the entries of $\tilde{B}$ and those of the upper triangular part of A_{11}. The other blocks of A, i.e., A_{12}, A_{21}, and A_{22} are derived as in [22]. This completes the proof for the case of identical singular values. The general case follows as in [22].

(2) $\Rightarrow$ (1) This is a straightforward modification of Theorem 7.1 of [22]. □

The canonical form presented in the previous theorem is determined both by discrete and continuous parameters. Of the discrete parameters, $n_1,\cdots,n_k$ are of particular importance. They indicate the multiplicities of the singular values $\sigma_1,\cdots,\sigma_k$ and determine the partitioning of the state-space systems into blocks of sizes $n_1,\cdots,n_k$. To each such block corresponds the n_j-dimensional subsystem (A_{jj}, B_j, C_j, D), $1\leqq j\leqq k$. Such a subsystem is a system with identical singular values σ_j. An interesting aspect of the canonical form is that the off-diagonal blocks A_{ij}, $i\neq j$, of the A matrix, which interconnect the various subsystems, are completely determined by the parameters of the diagonal subsystems. It therefore becomes clear that each system is made up of building blocks that are systems with identical singular values. The derivation of a canonical form, therefore, essentially reduces to the derivation of a canonical form for systems with identical singular values. The complexity of the canonical form depends crucially on min (m,p), the minimum of the dimensions of the input and output spaces. As a consequence, the canonical form, if specialized to single-input single-output systems, is considerably simplified.

The following corollary states the canonical form for single-input single-output systems.

COROLLARY 2.1. *The following two statements are equivalent*:

(1) $g(s) \in TC_n^{1,1}$.

(2) $g(s)$ *has a realization* $(A, b, c, d) \in \Re^{n \times n} \times \Re^{n \times 1} \times \Re^{1 \times n} \times \Re^{1 \times 1}$ *given by the parameters*

$$
\begin{array}{ll}
\sigma_1 > \cdots > \sigma_j > \cdots > \sigma_k > 0, & \\
n_1, \cdots, n_j, \cdots, n_k, & n_j \in \mathcal{N}, \quad \sum_{j=1}^{k} n_j = n; \\
s_1, \cdots, s_j, \cdots, s_k, & s_j = \pm 1, \quad 1 \leqq j \leqq k; \\
b_1, \alpha(1)_1, \cdots, \alpha(1)_j, \cdots, \alpha(1)_{n_1-1}, & b_1 > 0, \quad \alpha(1)_j > 0, \quad 1 \leqq j \leqq n_1 - 1; \\
\vdots & \\
b_i, \alpha(i)_1, \cdots, \alpha(i)_j, \cdots, \alpha(i)_{n_i-1}, & b_i > 0, \quad \alpha(i)_j > 0, \quad 1 \leqq j \leqq n_i - 1; \\
\vdots & \\
b_k, \alpha(k)_1, \cdots, \alpha(k)_j, \cdots, \alpha(k)_{n_k-1}, & b_k > 0, \quad \alpha(k)_j > 0, \quad 1 \leqq j \leqq n_k - 1; \\
d, & d \in \Re;
\end{array}
$$

in the following way:

(i) $b = (\underbrace{b_1, 0, \cdots, 0}_{n_1}, \cdots, \underbrace{b_j, 0, \cdots, 0}_{n_j}, \cdots, \underbrace{b_k, 0, \cdots, 0}_{n_k})^T$,

(ii) $c = (\underbrace{s_1 b_1, 0, \cdots, 0}_{n_1}, \cdots, \underbrace{s_j b_j, 0, \cdots, 0}_{n_j}, \cdots, \underbrace{s_k b_k, 0, \cdots, 0}_{n_k})$,

(iii) *For* $A =: (A_{ij})_{1 \leqq i,j \leqq k}$, $A_{ij} \in \Re^{n_i \times n_j}$, $1 \leqq i, j \leqq k$, *we have*

(a) *block diagonal entries* A_{jj}, $1 \leqq j \leqq k$:

$$
A_{jj} = \begin{pmatrix}
a_{jj} & \alpha(j)_1 & & & & \\
-\alpha(j)_1 & 0 & \alpha(j)_2 & & & \\
 & -\alpha(j)_2 & 0 & \cdot & & 0 \\
 & & \ddots & \ddots & \ddots & \\
 & 0 & & \cdot & 0 & \alpha(j)_{n_j-1} \\
 & & & & -\alpha(j)_{n_j-1} & 0
\end{pmatrix}
$$

with $a_{jj} = -b_j^2 / 2\sigma_j$.

(b) *off-diagonal blocks* A_{ij}, $1 \leqq i, j \leqq k$, $i \neq j$:

$$
A_{ij} = \begin{pmatrix}
a_{ij} & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & & \vdots \\
0 & 0 & \cdots & 0
\end{pmatrix} \quad \text{with } a_{ij} = \frac{-b_i b_j}{s_i s_j \sigma_i + \sigma_j},
$$

(iv) $d \in \Re$.

Moreover, (A, b, c, d) *as defined in* (2) *is balanced with gramian*

$$\Sigma = \operatorname{diag}(\sigma_1 I_{n_1}, \cdots, \sigma_j I_{n_j}, \cdots, \sigma_k I_{n_k}).$$

The map Γ, *which assigns to each system in* $C_n^{1,1}$ *the realization given in* (2), *is a canonical form.*

Remark 2.1. The canonical form of Theorem 2.1 is closely related to the canonical form derived in [22]. The main difference between the two canonical forms is in the structure of the B-matrix. In [22] a subblock B_j corresponding to a set of repeated singular values is constrained to have orthogonal rows. The implication of this is that the corresponding subblock of A has a desirable symmetry property. Here we do not impose this symmetry on A and hence we can relax the constraints on B_j and obtain a much simplified parametrization of B. Another advantage of this canonical form is that the parameters enter freely and are not constrained by the orthogonality assumption on the rows of B_j. Note, however, that there are no differences between the two canonical forms for the case of SISO systems and for multivariable systems if the singular values are distinct.

As pointed out above, the canonical form is essentially determined by the diagonal subsystems that are determined by the block partitioning of the system corresponding to the multiplicities of the singular values. It is therefore not surprising that the canonical form reduces considerably in complexity, especially in the multivariable case, if the system has distinct singular values.

The significance of the previous theorem and corollary lies in the fact that not only is it shown that each asymptotically stable system has a unique representation and a canonical form having a certain structure, but possibly of greater importance, it is shown that the converse is also true. If we have given an arbitrary set of parameters that satisfy the stated constraints and if a system is formed from these parameters, then the theorem guarantees that this system is automatically minimal and asymptotically stable. We therefore have a parametrization of the set of asymptotically stable systems of fixed dimension. Equivalently, we can interpret the theorem as providing a parametrization of the set of transfer functions of fixed McMillan degree whose poles are in the left halfplane.

Note that especially in the single-input single-output case the parameter space has a nice geometric structure, since the continuous parameters are only determined by simple inequality constraints.

An important feature of a balanced realization of a linear system is its close relationship to the Hankel operator corresponding to the system. We define an *integral Hankel operator* with kernel $H(t) \in \mathfrak{R}^{p\times m}$, $t \geqq 0$, given by

$$\mathcal{H}: L^2_{\mathfrak{R}^m}([0,\infty[) \to L^2_{\mathfrak{R}^p}([0,\infty[),$$

$$u(t) \mapsto (\mathcal{H}(u))(s) = \int_0^\infty H(t+s)u(t)\,dt.$$

We assume that $\mathcal{H}$ is well defined and a finite rank operator. The singular values $(\sigma_j)_{1\leqq j \leqq n}$ of $\mathcal{H}$ are defined to be the nonzero eigenvalues of $(\mathcal{H}^*\mathcal{H})^{1/2}$ ordered with respect to their magnitude and taking into account their multiplicities. Under these conditions there exist families of orthonormal vectors (Schmidt vectors) $(v_j)_{1\leqq j\leqq n}$ and $(w_j)_{1\leqq j \leqq n}$ in $L^2_{\mathfrak{R}^m}([0,\infty[)$ and $L^2_{\mathfrak{R}^p}([0,\infty[)$, respectively, such that

$$\mathcal{H}v_j = \sigma_j w_j, \quad \mathcal{H}^* w_j = \sigma_j v_j, \quad 1 \leqq j \leqq n.$$

The significance of Hankel operators in a system-theoretic context is that for a system $(A, B, C, D) \in C_n^{p,m}$ the Hankel operator $\mathcal{H}$ with kernel $H(t) := Ce^{tA}B$ can be interpreted as an operator mapping past inputs to future outputs. If (A, B, C, D) is balanced, it can easily be verified that the singular values of the Hankel operator $\mathcal{H}$ equal the singular values of the system (A, B, C, D) [10]. Moreover, $v_j(t) = B^Te^{tA^T}e_j/\sqrt{\sigma_j}$ and $w_j(t) = Ce^{tA}e_j/\sqrt{\sigma_j}$, $1 \leqq j \leqq n$. Since $v_j(0) = B^Te_j/\sqrt{\sigma_j}$ and $w_j(0) = Ce_j/\sqrt{\sigma_j}$, $1 \leqq j \leqq n$, the starting points of the Schmidt vectors are fully determined by the B and C matrices and the singular values. This gives an interpretation of some of the parameters of a system in canonical form in terms of analytical properties of the corresponding Hankel operator.

By standard realization theory we know that there is a one-to-one correspondence between Hankel operators of rank n and asymptotically systems of McMillan degree n. The discussion in the previous paragraph implies that finding a canonical form for linear systems in terms of balanced realizations is equivalent to defining a unique basis for the eigenspaces of the nonzero eigenvalues of the operator $\mathscr{H}^*\mathscr{H}$. This observation gives another interpretation of the complexity of the canonical form in the case of repeated singular values.

If we now consider the Hankel operator $\mathscr{H}$ corresponding to a scalar transfer function in $TC_n^{1,1}$, the eigenvectors and Schmidt vectors of $\mathscr{H}$ coincide (up to a sign), since $\mathscr{H}$ is self-adjoint. If, moreover, (A, b, c, d) is in the canonical form of the previous corollary, the eigenvectors, respectively, Schmidt vectors, are given by $v_j(t) = b^T e^{tA^T} e_j/\sqrt{\sigma_j} = \tilde{s}_j w_j(t) = (\tilde{s}_j/\sqrt{\sigma_j}) c e^{tA} e_j$, $1 \leqq j \leqq n$, where we have used that $A = SA^TS$ and $c^T = Sb$ for $S = \operatorname{diag}(\tilde{s}_1, \cdots, \tilde{s}_j, \cdots, \tilde{s}_n) := \operatorname{diag}(s_1 \hat{I}_{n_1}, \cdots, s_j \hat{I}_n, \cdots, s_k \hat{I}_{n_k})$.

The eigenvalue corresponding to v_j is therefore calculated to be $\tilde{s}_j\sigma$, since

$$(\mathscr{H}v_j)(t) = \int_0^\infty H(t+s) v_j(s)\, ds = \frac{1}{\sqrt{\sigma_j}} \int_0^\infty c e^{(t+s)A} b b^T e^{sA^T} e_j\, ds$$

$$= \sigma_j \frac{1}{\sqrt{\sigma_j}} c e^{tA} e_j = \tilde{s}_j \sigma_j v_j(t).$$

This observation allows us to use the canonical form for scalar systems to investigate the dimensions of the eigenspaces of a finite rank Hankel operator.

THEOREM 2.2. *Let $h(t) \in \mathfrak{R}$, $t \geqq 0$, be the kernel of a finite rank Hankel operator $\mathscr{H}$ acting on $L^2_{\mathfrak{R}}([0, \infty[)$. If λ is a nonzero eigenvalue of $\mathscr{H}$, then*

$$|\dim(\ker(\lambda I - \mathscr{H})) - \dim(\ker(-\lambda I - \mathscr{H}))| \leqq 1.$$

Proof. By standard realization theory, $h(t)$, $t \geqq 0$, has a realization $(A, b, c, d) \in C_n^{1,1}$, where $n = \operatorname{rank}(\mathscr{H})$, i.e., $h(t) = c e^{tA} b$, $t \geqq 0$. We can assume that (A, b, c, d) is in the canonical form of Corollary 2.1. All nonzero eigenvalues of $\mathscr{H}$ are given by the diagonal entries of $S\Sigma = \operatorname{diag}(s_1\sigma_1 \hat{I}_{n_1}, \cdots, s_j\sigma_j \hat{I}_{n_j}, \cdots, s_k\sigma_k \hat{I}_{n_k})$. We know that $|\lambda| = |s_{i_0}\sigma_{i_0}|$ for some $1 \leqq i_0 \leqq k$, and therefore

$$|\dim(\ker(\lambda I - \mathscr{H})) - \dim(\ker(-\lambda I - \mathscr{H}))| = \operatorname{trace}|(s_{i_0} \hat{I}_{n_{i_0}})| \leqq 1. \qquad \square$$

As a simple corollary to this theorem we have that Hankel operators with positive eigenvalues cannot have repeated singular values.

COROLLARY 2.2. *If $\mathscr{H}$ as defined in the theorem has only nonnegative eigenvalues, then the multiplicity of each of the singular values is one.*

Note that single-input single-output systems whose corresponding Hankel operators have only positive nonzero eigenvalues can be characterized as relaxation systems [25].

Remark 2.2. Using a canonical form for symmetric multivariable systems, a similar result was obtained in [23] for self-adjoint Hankel operators acting on vector-valued spaces. More precisely, if $\mathscr{H}$ is a finite rank, self-adjoint Hankel operator acting on the space $L^2_{\mathfrak{R}^p}([0, \infty[)$, then

$$|\dim(\ker(\lambda I - \mathscr{H})) - \dim(\ker(-\lambda I - \mathscr{H}))| \leqq p,$$

where λ is a nonzero eigenvalue of $\mathscr{H}$.

The following remark gives further interpretations of the parameters of the canonical form and relates them to important analytical properties of the system.

Remark 2.3. If $(A, b, c, d) \in C_n^{1,1}$ is in balanced canonical form with $g(s) = c(sI - A)^{-1}b$ and $h(t) = ce^{tA}b$, $t \geqq 0$, then we have the following properties:

(i) $\|h(t)\|_2^2 := \int_0^\infty h(t)^2\,dt = \sum_{j=1}^k \sigma_j b_j^2$,

(ii) $\|g(s)\|_\infty := \sup_{w \in \Re} |g(iw)| \leqq 2\sum_{j=1}^k \sigma_j$,

(iii) $g(0) = \int_0^\infty h(t)\,dt = 2\sum_{j=1}^k s_j\sigma_j(\sum_{i=1}^{n(j)}(-1)^{i+1})$.

Note that if (A, b, c, d) has distinct singular values and $s_j = 1$ or $s_j = -1$, for all $1 \leqq j \leqq n$, then (ii) and (iii) imply that the bound in (ii) is attained, i.e.,

$$\|g(s)\|_\infty = 2\sum_{j=1}^n \sigma_j.$$

References to these results, which were slightly adapted here to our particular canonical form, can be found in [18] for (i) and (iii), [33] for (iii), and [10] for (ii).

3. Allpass systems. In the previous section we have seen that one of the main structural elements of the canonical form for asymptotically stable systems are systems with identical singular values. In this section we give an interpretation of such systems.

It was shown in [10] that each strictly proper system with identical singular values is the strictly proper part of an allpass system. Conversely, each allpass system has identical singular values. We can therefore say that the "building blocks" of the general canonical form are the strictly proper parts of allpass systems.

DEFINITION 3.1. A system $(A, B, C, D) \in C_n^{m,m}$ is called *allpass* if for $G(s) = C(sI - A)^{-1}B + D$ we have

$$G(s)G(-s)^T = \sigma^2 I, \qquad s \in \mathscr{C},$$

for some $\sigma > 0$. We denote by A_n^m the subset of $C_n^{m,m}$ containing all allpass systems. The set of transfer functions of systems in A_n^m is denoted by TA_n^m.

Remark 3.1. The usual definition of an allpass system is generalized here to the case where σ is not necessarily equal to one. Note that in the mathematical literature the transfer functions of allpass systems are referred to as inner functions in the case of $\sigma = 1$.

Allpass systems or inner functions play an important role in many aspects of control theory, circuit theory, and mathematics. It is therefore of interest to have a canonical form and a parametrization of these systems. To obtain such a canonical form and parametrization of allpass systems from the results for systems with identical singular values, it is necessary to impose a relation between the parameters U, $\tilde{B}$, D, and σ.

A canonical form for allpass systems is given in the following theorem.

THEOREM 3.1. *The following two statements are equivalent*:

(1) $G(s) \in TA_n^m$.

(2) $G(s)$ *has a realization* $(A, B, C, D) \in \Re^{n\times n} \times \Re^{n\times m} \times \Re^{m\times n} \times \Re^{m\times m}$ *given by the parameters*: $\sigma > 0$; $r \in \mathscr{N}$, $1 \leqq r \leqq n$; $\tilde{B} \in \Re^{r\times m}$ *is positive upper triangular*; $\tilde{A}$ *is in r-balanced form*; $D \in \Re^{m\times m}$ *is such that* $DD^T = \sigma^2 I$; $U = -D\tilde{B}^T\Delta^{-1}(\sqrt{\sigma})^{-1}$ *with* $\Delta = (\tilde{B}\tilde{B}^T)^{1/2}$; *in the following way*:

(i) $B = \begin{pmatrix} \tilde{B} \\ 0 \end{pmatrix}$;

(ii) $C = (U\Delta, 0)$ *where* $\Delta = (\tilde{B}\tilde{B}^T)^{1/2}$;

(iii) $A = \tilde{A} - \frac{1}{\sigma}[\operatorname{diag}(\Delta^2, 0)]_l - \frac{1}{2\sigma}[\operatorname{diag}(\Delta^2, 0)]_d$;

(iv) $D \in \Re^{m \times m}$ *is such that* $DD^T = \sigma^2 I$.

Moreover, (A, B, C, D) *as defined in* (2) *is balanced with gramian* $\Sigma = \sigma I_n$. *The map* Γ_a, *which assigns to each system in* A_n^m *the realization given in* (2), *is a canonical form.*

Proof. The proof is an application of Theorem 5.1 of [10] and Theorem 2.1. □

It follows from the previous theorem that a parametrization of systems with identical singular values immediately leads to a parametrization of allpass systems and vice versa. For SISO systems this canonical form has a particularly simple structure.

COROLLARY 3.1. *The following two statements are equivalent:*

(1) $g(s) \in TA_n^1$.

(2) $g(s)$ *has a realization* $(A, b, c, d) \in \Re^{n \times n} \times \Re^{n \times 1} \times \Re^{1 \times n} \times \Re^{1 \times 1}$ *given by the parameters*

$$\sigma > 0, \quad b > 0, \quad \alpha_j > 0, \quad 1 \leqq j \leqq n-1, \quad s_1 = \pm 1;$$

in the following way:

(i) $b = (b, 0, \cdots, 0)^T$;

(ii) $c = (s_1 b, 0, \cdots, 0)$;

(iii) $$A = \begin{pmatrix} a & \alpha_1 & & & & \\ -\alpha_1 & 0 & \alpha_2 & & & \\ & -\alpha_2 & 0 & \cdot & & 0 \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \cdot & 0 & \alpha_{n-1} \\ & & & & -\alpha_{n-1} & 0 \end{pmatrix} \quad \textit{with } a = \frac{-b^2}{2\sigma};$$

(iv) $d = -s_1 \sigma$.

Moreover, (A, b, c, d) *as defined in* (2) *is balanced with gramian* $\Sigma = \sigma I_n$. *The map* Γ_a *which assigns to each system in* A_n^1 *the realization given in* (2), *is a canonical form.* □

Remark 3.2. The A-matrix in the canonical form for SISO allpass transfer functions is closely related to the so-called Schwarz form for matrices, which has been studied in connection with stability tests for matrices (see, e.g., [4]). The recursive structure of the tridiagonal matrix permits us to give explicit realization algorithms for allpass systems [21]. It also follows easily from this recursive structure that each allpass transfer function $g(s)$ can be written as a continued fraction as follows:

$$g(s) = -s_1\sigma + \cfrac{s_1 b^2}{s + \cfrac{b^2}{2\sigma} + \cfrac{\alpha_1^2}{s + \cfrac{\alpha_2^2}{\ddots \, s + \cfrac{\cdot}{s + \cfrac{\alpha_{n-1}^2}{s}}}}},$$

where $s_1, b, \sigma, \alpha_1, \cdots, \alpha_{n-1}$ are the same parameters as in the canonical form of the corollary.

4. Minimal systems. To apply the balancing technique in § 2, a system must be asymptotically stable. For minimal systems that are not necessarily asymptotically stable, Jonckheere and Silverman [16] have introduced a method that is not based on balancing solutions to Lyapunov equations but is instead based on balancing solutions to Riccati equations. Their definition of a Riccati-balanced system was extended in [27] to include a feed-through term.

DEFINITION 4.1. The class of systems $(A, B, C, D) \in \Re^{n\times n}\times\Re^{n\times m}\times\Re^{p\times n}\times\Re^{p\times m}$, which are minimal, is denoted by $L_n^{p,m}$. A system $(A, B, C, D)\in L_n^{p,m}$ is called *Riccati balanced* if there exists a positive diagonal matrix $\Sigma_l = \operatorname{diag}(q_1, \cdots, q_j, \cdots, q_n)$ such that

$$\begin{aligned}(A-BS_r^{-1}D^TC)^T\Sigma_l+\Sigma_l(A-BS_r^{-1}D^TC)-\Sigma_lBS_r^{-1}B^T\Sigma_l+C^TR_r^{-1}C=0,\\ (A-BS_r^{-1}D^TC)\Sigma_l+\Sigma_l(A-BS_r^{-1}D^TC)^T-\Sigma_lC^TR_r^{-1}C\Sigma_l+BS_r^{-1}B^T=0,\end{aligned}\tag{2}$$

where $S_r = I+D^TD$ and $R_r = I+DD^T$. Σ_l is called the *Riccati gramian* of the system.

The following canonical form for systems in $L_n^{p,m}$ is a modification of a canonical form derived in [27]. The canonical form presented here differs in two ways from the canonical form in [27]. The parametrization of the B-matrix has been changed analogously to the changes for the canonical form for asymptotically stable systems. Whereas in the present paper the B-matrix is chosen as a parameter of the system, in [27] the C-matrix serves as a parameter. As in the case of the present canonical form, such a change of parameter can be performed for all the canonical forms presented in this paper. The calculations involved are tedious but not difficult if it is noted that for $D\in\Re^{p\times m}$ we have that $D^T(I+DD^T)^{1/2}=(I+D^TD)^{1/2}D^T$.

THEOREM 4.1. *The following two statements are equivalent*:

(1) $G(s)\in TL_n^{p,m}$.

(2) *$G(s)$ has a realization $(A, B, C, D)\in\Re^{n\times n}\times\Re^{n\times m}\times\Re^{p\times n}\times\Re^{p\times m}$ given by the parameters*

$$\begin{array}{ll} q_1>\cdots>q_j>\cdots>q_k>0, & \\ n_1,\cdots,n_j,\cdots,n_k, & n_j\in\mathcal{N},\quad \sum_{j=1}^k n_j=n;\\ r_1,\cdots,r_j,\cdots,r_k, & r_j\in\mathcal{N},\quad 1\leqq r_j\leqq\min(n_j,m,p);\\ U_1,\cdots,U_j,\cdots,U_k, & U_j\in\Re^{p\times r_j},\quad U_j^TU_j=I_{r_j};\\ \tilde B_1,\cdots,\tilde B_j,\cdots,\tilde B_k, & \tilde B_j\in\Re^{r_j\times m}\ \textit{positive upper triangular};\\ \tilde A_1,\cdots,\tilde A_j,\cdots,\tilde A_k, & \tilde A_j\in\Re^{n_j\times n_j}\ \textit{in } r_j\textit{-balanced form};\\ D, & D\in\Re^{p\times m};\end{array}$$

in the following way:

If (A, B, C, D) is partitioned according to $n_1,\cdots,n_j,\cdots,n_k$, then

(i) $B_j=\begin{pmatrix}\tilde B_jS_r^{1/2}\\0\end{pmatrix}$ *where* $S_r=I+D^TD,\quad 1\leqq j\leqq k;$

(ii) $C_j=(R_r^{1/2}U_j\Delta_j,0)$ *where* $R_r=I+DD^T,\quad \Delta_j=(\tilde B_j\tilde B_j^T)^{1/2},\quad 1\leqq j\leqq k;$

(iii) $$A_{jj}=\tilde A_j-\frac{1-q_j^2}{q_j}[\operatorname{diag}(\Delta_j^2,0)]_l-\frac{1-q_j^2}{2q_j}[\operatorname{diag}(\Delta_j^2,0)]_d+\operatorname{diag}(\tilde B_jD^TU_j\Delta_j,0),\qquad 1\leqq j\leqq k;$$

(iv) $$A_{ij}=\frac{1}{q_i^2-q_j^2}\operatorname{diag}\left(\left(q_j(1+q_i^2)\tilde{B}_i\tilde{B}_j^T-q_i(1+q_j^2)\Delta_iU_i^TU_j\Delta_j\right),0\right)$$
$$+\operatorname{diag}(\tilde{B}_iD^TU_j\Delta_j,0),\qquad 1\leqq i,j\leqq k,\quad i\neq j;$$

(v) $D\in\Re^{p\times m}$.

Moreover, (A, B, C, D) as defined in (2), is Riccati balanced with gramian

$$\Sigma_l=\operatorname{diag}(q_1I_{n_1},\cdots,q_jI_{n_j},\cdots,q_kI_{n_k}).$$

The map Γ_l, which assigns to each system in $L_n^{p,m}$ the realization given in (2), is a canonical form. □

This canonical form reduces as follows to scalar systems.

COROLLARY 4.1. *The following two statements are equivalent*:

(1) $g(s)\in TL_n^{1,1}$.

(2) *$g(s)$ has a realization $(A, b, c, d)\in\Re^{n\times n}\times\Re^{n\times 1}\times\Re^{1\times n}\times\Re^{1\times 1}$ given by the parameters*

$$\begin{array}{rllll}
q_1>\cdots>q_j>\cdots>q_k>0, & & & & \\
n_1,\cdots,n_j,\cdots,n_k, & n_j\in\mathcal{N}, & \sum_{j=1}^k n_j=n; & & \\
s_1,\cdots,s_j,\cdots,s_k, & s_j=\pm1, & 1\leqq j\leqq k; & & \\
b_1,,\alpha(1)_1,\cdots,\alpha(1)_j,\cdots,\alpha(1)_{n_1-1}, & b_1>0, & \alpha(1)_j>0, & 1\leqq j\leqq n_1-1; \\
\vdots & & & \\
b_i,\alpha(i)_1,\cdots,\alpha(i)_j,\cdots,\alpha(i)_{n_i-1}, & b_i>0, & \alpha(i)_j>0, & 1\leqq j\leqq n_i-1; \\
\vdots & & & \\
b_k,\alpha(k)_1,\cdots,\alpha(k)_j,\cdots,\alpha(k)_{n_k-1}, & b_k>0, & \alpha(k)_j>0, & 1\leqq j\leqq n_k-1; \\
d, & d\in\Re; & &
\end{array}$$

in the following way:

(i) $$b=(\underbrace{b_1,0,\cdots,0}_{n_1},\cdots,\underbrace{b_j,0,\cdots,0}_{n_j},\cdots,\underbrace{b_k,0,\cdots,0}_{n_k})^T,$$

(ii) $$c=(\underbrace{s_1b_1,0,\cdots,0}_{n_1},\cdots,\underbrace{s_jb_j,0,\cdots,0}_{n_j},\cdots,\underbrace{s_kb_k,0,\cdots,0}_{n_k}),$$

(iii) *For $A=:(A_{ij})_{1\leqq i,j\leqq k}$ we have*

(a) *block diagonal entries A_{jj}, $1\leqq j\leqq k$*:

$$A_{jj}=\begin{pmatrix}
a_{jj} & \alpha(j)_1 & & & & \\
-\alpha(j)_1 & 0 & \alpha(j)_2 & & & \\
 & -\alpha(j)_2 & 0 & \cdot & & 0 \\
 & & \ddots & \ddots & \ddots & \\
 & 0 & & \cdot & 0 & \alpha(j)_{n_j-1} \\
 & & & & -\alpha(j)_{n_j-1} & 0
\end{pmatrix}$$

with $$a_{jj}=\frac{-b_j^2}{1+d^2}\left(\frac{1-q_j^2}{2q_j}-s_jd\right);$$

(b) *off-diagonal blocks* A_{ij}, $1 \leqq i, j \leqq k$, $i \neq j$:

$$A_{ij} = \begin{pmatrix} a_{ij} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \quad \textit{with } a_{ij} = \frac{-b_i b_j}{1+d^2}\left(\frac{1 - s_i s_j q_i q_j}{s_i s_j q_i + q_j} - s_j d\right);$$

(iv) $d \in \mathfrak{R}$.

Moreover, (A, b, c, d) *as defined in* (2) *is Riccati balanced with Riccati gramian*

$$\Sigma_l = \operatorname{diag}(q_1 I_{n_1}, \cdots, q_j I_{n_j}, \cdots, q_k I_{n_k}).$$

The map Γ_l, *which assigns to each system in* $L_n^{1,1}$ *the realization given in* (2), *is a canonical form.*

Proof. The corollary follows immediately from the theorem and a straightforward reparametrization of the entries of the b-vector. □

It is instructive to note that this canonical form for minimal systems and the corresponding parametrization has essentially the same structure as the canonical form for asymptotically stable minimal systems. In fact, the only difference is the way in which the parameters enter those entries of the A-matrix that are functions of the other parameters. In later sections we will find that the same applies to all classes of systems considered in this paper. It should be noted, however, that in general it is not clear how the parameters of an asymptotically stable system that is parametrized using the above canonical form are related to those parameters with which the system is parametrized in the canonical form for asymptotically stable systems.

The way the canonical form for minimal systems was derived in [27] was by relating a minimal system to the state-space realization of its normalized left coprime factorization. If $G(s)$ is the transfer function of a linear system, then the normalized left coprime factorization $[N(s), M(s)]$ of $G(s)$ is defined such that $G(s) = M(s)^{-1}N(s)$, where $N(s)$, $M(s)$ are asymptotically stable with $M(s)$ proper and satisfy $NN^* + MM^* = I$. Such coprime factorizations play an important role in modern control theory [31]. If a state-space realization of $G(s)$ is available, a state-space realization of $[N(s), M(s)]$ can be calculated by solving the Riccati equations corresponding to the state-space realization of $G(s)$. The canonical form for minimal systems can then be derived by calculating the canonical form for asymptotically stable systems of the asymptotically stable coprime factors. Exploiting the state-space formulae that relate a state-space realization of a system to the state-space realization of its coprime factors, it was possible to derive the canonical form and the parametrization result for minimal systems.

In the following sections we will use the same approach to the derivation of canonical forms for other classes of systems. Rather than deriving a canonical form separately for each class of systems, explicit maps will be constructed to relate the state-space realizations of the class of systems to an appropriate subclass of the set of asymptotically stable systems. Thereby the canonical form for asymptotically stable systems can be exploited to derive a canonical form for other classes of systems.

5. Bounded real systems. An important subclass of asymptotically stable systems is that with transfer functions bounded by one on the imaginary axis. This class of so-called bounded real systems is used to parametrize all stabilizing controllers of a plant such that the closed-loop system satisfies an H^∞ constraint (see, e.g., [11]). In fact, transfer functions of bounded real systems are the real rational functions in the open unit ball of H^∞.

DEFINITION 5.1. Let $(A, B, C, D) \in C_n^{p,m}$ be such that $I - D^T D > 0$; then (A, B, C, D) is called *bounded real* if for $G(s) = C(sI - A)^{-1}B + D$ we have

$$I - G(-iw)^T G(iw) > 0, \qquad w \in \Re.$$

We denote by $B_n^{p,m}$ the subset of $C_n^{p,m}$ containing all bounded real systems. $TB_n^{p,m}$ denotes the set of transfer functions of systems in $B_n^{p,m}$.

Remark 5.1. Other authors call an asymptotically stable system bounded real if

$$I - G(-iw)^T G(iw) \geqq 0, \qquad w \in \Re.$$

In this section we will derive a canonical form for bounded real systems. To do this we first must define what we mean by balancing of bounded real systems. With the classes of systems that we considered in the previous sections, we associated certain Lyapunov and Riccati equations. A system was called balanced if the solutions to the corresponding equations were balanced, i.e., equal and diagonal. Bounded real systems can be shown to satisfy the so-called bounded real Riccati equation. We proceed analogously to define a balanced realization for bounded real systems. The following proposition states several well-known results relating bounded real systems to this Riccati equation (see, e.g., [5], [32]).

PROPOSITION 5.1. *Let $(A, B, C, D) \in C_n^{p,m}$ such that $S := I - D^T D > 0$ and $G(s) = C(sI - A)^{-1}B + D$; then*

(1) $I - G(-iw)^T G(iw) \geqq 0$, $w \in \Re$, *if and only if there exists $P = P^T > 0$ such that*

$$A^T P + PA + C^T C + (PB + C^T D)S^{-1}(PB + C^T D)^T = 0. \tag{3}$$

We call this Riccati equation the bounded real Riccati equation (BRRE).

(2) *If either of the conditions in* (1) *is satisfied, then $P = P^T > 0$ for any solution to* (3). *There exist solutions $P_{\min}$ and $P_{\max}$ to* (3) *such that for any solution $P = P^T$ we have*

$$0 < P_{\min} \leqq P \leqq P_{\max}.$$

(3) *If (A, B, C, D) is bounded real, i.e., $I - G(-iw)^T G(iw) > 0$, $w \in \Re$, then*

$$0 < P_{\min} < P_{\max}$$

and $P_{\min}$ is the unique solution to (3) *such that $A + BS^{-1}(B^T P + D^T C)$ is asymptotically stable.*

(4) *(A, B, C, D) is bounded real if there exists a solution $P = P^T > 0$ to* (3) *such that $A + BS^{-1}(B^T P + D^T C)$ is asymptotically stable. Moreover, $P_{\min}$ is the unique such solution.*

(5) *A system (A, B, C, D) is bounded real if and only if its dual system (A^T, C^T, B^T, D^T) is bounded real. If (A, B, C, D) is bounded real with $P_{\min}$ and $P_{\max}$ the minimal, respectively, maximal solution to* (3), *then $P_{\min}^{-1}$ is the maximal and $P_{\max}^{-1}$ the minimal solution to the BRRE corresponding to (A^T, C^T, B^T, D^T).*

A problem with the bounded real Riccati equation is that there is no unique positive-definite solution. Opdenacker and Jonckheere [28] introduced a balancing technique for bounded real systems by balancing the minimal solution with the inverse of the maximal solution to the BRRE. We will follow their definition since it is possible to derive the desired canonical form and realization result using this particular choice of solution. Note that balancing the minimal solution of the BRRE with the inverse of its maximal solution is the same as balancing the minimal solution with the minimal solution of the dual equation.

DEFINITION 5.2. A system $(A, B, C, D) \in B_n^{p,m}$ is called *bounded real balanced* if

$$P_{\min} = P_{\max}^{-1} = \operatorname{diag}(p_1, \cdots, p_j, \cdots, p_n) =: \Sigma_b,$$

where $P_{\min}$, $P_{\max}$ are the minimal, respectively, maximal solution to the BRRE. Σ_b is called the *bounded real gramian* of (A, B, C, D).

Remark 5.2. Note that since $P_{\max} > P_{\min}$ for systems in $B_n^{p,m}$ we have that $0 < \Sigma_b < I$.

Before we can derive a canonical form for bounded real systems we need several lemmas. The first of these states standard identities that we will frequently need.

LEMMA 5.1. *If $D \in \Re^{p\times m}$ is such that $S := I - D^TD > 0$ and $R := I - DD^T > 0$, then S and R have the following properties*:

(i) $D^TR^{-1} = S^{-1}D^T$,

(ii) $D^TR^{1/2} = S^{1/2}D^T$.

The canonical form for bounded real systems will be derived by mapping bounded real systems to a certain class of asymptotically stable systems. The significance of this map is that it maps bounded real balanced systems to asymptotically stable systems that are balanced with respect to the corresponding Lyapunov equations. A canonical form for bounded real systems can therefore be derived by bringing the associated asymptotically stable systems to the canonical form of Theorem 2.1. Reversing the map, we can obtain the desired canonical form for bounded real systems.

The following lemma establishes the map between bounded real systems and a subclass of asymptotically stable systems.

LEMMA 5.2. *Let $(A, B, C, D) \in B_n^{p,m}$. If $P_{\min}$ is the minimal and $P_{\max}$ the maximal solution to the* BRRE, *then with $S = I - D^TD$ and $R = I - DD^T$ we have that*

$$(A_c, B_c, C_c, D_c) := \left(A + BS^{-1}D^TC, [BS^{-1/2}, P_{\max}^{-1}C^TR^{-1/2}], \begin{bmatrix} R^{-1/2}C \\ S^{-1/2}B^TP_{\min} \end{bmatrix}, D\right)$$

$$\in C_n^{p+m,p+m}$$

and

$$A_cP_{\max}^{-1} + P_{\max}^{-1}A_c^T = -B_cB_c^T, \tag{4}$$

$$A_c^TP_{\min} + P_{\min}A_c = -C_c^TC_c. \tag{5}$$

Proof. First note that $\tilde{P} = \tilde{P}^T > 0$ solves the bounded real Riccati equation

$$A^TP + PA + C^TC + (PB + C^TD)S^{-1}(PB + C^TD)^T = 0$$

if and only if $\tilde{P}^{-1}$ solves the dual equation

$$AP + PA^T + BB^T + (PC^T + BD^T)R^{-1}(PC^T + BD^T)^T = 0.$$

It is now easy to verify that these two Riccati equations can be rewritten as the Lyapunov equations (4), (5) with $P_{\max}^{-1}$ and $P_{\min}$ as their solutions. Since $P_{\max}^{-1} > 0$ and $P_{\min} > 0$ we have that (A_c, B_c, C_c, D_c) is minimal if and only if A_c is asymptotically stable. Since (A_c, B_c, D_c, C_c) satisfies the Lyapunov equations (4), (5) we know that the eigenvalues of A_c are in the closed left halfplane. To show that they are in fact in the open left halfplane, assume there exists $w \in \Re$ and $x \in \mathscr{C}^n$ such that

$$A_cx = iwx.$$

By applying x to the right and x^* to the left of (5), i.e.,

$$x^*A_c^TP_{\min}x + x^*P_{\min}A_cx = 0 = -x^*C_c^TC_cx,$$

we obtain that $Cx=0$ and thus

$$A_c x = (A + BS^{-1}D^TC)x = Ax = iwx.$$

Hence by the PBH test we have that (A, B, C, D) is not observable, which is a contradiction to the assumption. This implies the asymptotic stability of A_c and the minimality of (A_c, B_c, C_c, D_c). □

In the following lemma the inverse of the map of the previous lemma will be investigated.

LEMMA 5.3. *Let* $(A, B, C, D) \in \Re^{n\times n} \times \Re^{n\times m} \times \Re^{p\times n} \times \Re^{p\times m}$ *be such that* $S := I - D^TD > 0$, $R := I - DD^T$, *and* $P = P^T > 0$, *such that* $P^{-1} > P$. *If*

$$(A_c, B_c, C_c, D_c) := \left(A + BS^{-1}D^TC, [BS^{-1/2}, PC^TR^{-1/2}], \begin{bmatrix} R^{-1/2}C \\ S^{-1/2}B^TP \end{bmatrix}, D \right) \in C_n^{p+m,p+m}$$

with

$$A_cP + PA_c^T = -B_cB_c^T, \tag{6}$$

$$A_c^TP + PA_c = -C_c^TC_c, \tag{7}$$

then $(A, B, C, D) \in C_n^{p,m}$. *Under the same conditions the eigenvalues of* $A + BS^{-1}(B^TP + D^TC)$ *are in the open left halfplane.*

Proof. First note that the Lyapunov equation (6) can be rewritten as

$$\begin{aligned} 0 = {} & (A + BS^{-1}D^TC + BS^{-1}B^TP)P + P(A + BS^{-1}D^TC + BS^{-1}B^TP)^T \\ & + (I - P^2)BS^{-1}B^T(I - P^2) - P^2BS^{-1}B^TP^2 + PC^TR^{-1}CP, \end{aligned} \tag{8}$$

which is equivalent to

$$\begin{aligned} 0 = {} & (A + BS^{-1}D^TC + BS^{-1}B^TP)^TP^{-1} + P^{-1}(A + BS^{-1}D^TC + BS^{-1}B^TP) \\ & + (P^{-1} - P)BS^{-1}B^T(P^{-1} - P) - PBS^{-1}B^TP + C^TR^{-1}C. \end{aligned} \tag{9}$$

Similarly, (7) can be rewritten as

$$\begin{aligned} 0 = {} & (A + BS^{-1}D^TC + BS^{-1}B^TP)^TP + P(A + BS^{-1}D^TC + BS^{-1}B^TP) \\ & - PBS^{-1}B^TP + C^TR^{-1}. \end{aligned} \tag{10}$$

Subtracting (10) from (9) we obtain

$$\begin{aligned} 0 = {} & (A + BS^{-1}D^TC + BS^{-1}B^TP)^T(P^{-1} - P) \\ & + (P^{-1} - P)(A + BS^{-1}D^TC + BS^{-1}B^TP) + (P^{-1} - P)BS^{-1}B^T(P^{-1} - P). \end{aligned} \tag{11}$$

Since $P^{-1} - P > 0$ we can pre- and post-multiply this equation by $(P^{-1} - P)^{-1}$, and hence

$$\begin{aligned} 0 = {} & (A + BS^{-1}D^TC + BS^{-1}B^TP)(P^{-1} - P)^{-1} \\ & + (P^{-1} - P)^{-1}(A + BS^{-1}D^TC + BS^{-1}B^TP)^T + BS^{-1}B^T. \end{aligned} \tag{12}$$

Since $(P^{-1} - P)^{-1} > 0$ this Lyapunov equation implies that $A + BS^{-1}D^TC + BS^{-1}B^TP$ is asymptotically stable if and only if $(A + BS^{-1}D^TC + BS^{-1}B^TP, BS^{-1/2})$ is controllable. If we assume that this is not the case, then there exist $\lambda \in \mathscr{C}$ and $x \in \mathscr{C}^n$ such that

$$x^*(A + BS^{-1}D^TC + BS^{-1}B^TP) = \lambda x^*, \qquad x^*BS^{-1/2} = 0.$$

Pre- and post-multiplying (12) by x^* and x, respectively, we obtain that $\operatorname{Re}(\lambda) \leqq 0$. Now assume that $\lambda = iw$, $w \in \Re$; then,

$$x^*(A + BS^{-1}D^TC) = iwx^*,$$

which is a contradiction to the asymptotic stability of $A + BS^{-1}D^TC$. Hence we have the controllability of

$$(A + BS^{-1}D^TC + BS^{-1}B^TP, BS^{-1/2})$$

and the asymptotic stability of $A + BS^{-1}D^TC + BS^{-1}B^TP$. An application of the PBH test now implies the controllability of (A, B).

We can show similarly that

$$(A + BS^{-1}D^TC + PC^TR^{-1}C, R^{-1/2}C)$$

is observable and hence that (A, C) is observable. Having shown the minimality of (A, B, C, D) it remains to show that A is asymptotically stable.

Note that the Lyapunov equation (7) can be rewritten as

$$A^TP + PA + C^TC + (B^TP + D^TC)^TS^{-1}(B^TP + D^TC) = 0.$$

Thus A is asymptotically stable if and only if

$$\left(A, \begin{bmatrix} C \\ S^{-1/2}(B^TP + D^TC) \end{bmatrix}\right)$$

is observable. But if this system is not observable we have that (A, C) is not observable, which is a contradiction to the minimality of (A, B, C, D).

We are now in a position to prove the main theorem of this section, which establishes a canonical form for bounded real systems. Note that the main problem in the proof of the theorem will be to show the "parametrization part" of the result, i.e., that a system with a particular structure is indeed bounded real. Due to the particular parametrization of the systems, a solution of the bounded real Riccati equation can be written immediately. To show that the system is bounded real, it is, however, necessary to show that this solution is the minimal solution. In the previous lemma we have already gone some way toward showing this.

THEOREM 5.1. *The following two statements are equivalent*:

(1) $B(s) \in TB_n^{p,m}$.

(2) $B(s)$ *has a realization* $(A, B, C, D) \in \Re^{n\times n} \times \Re^{n\times m} \times \Re^{p\times n} \times \Re^{p\times m}$ *given by the parameters*:

$$\begin{array}{ll}
1 > p_1 > \cdots > p_j > \cdots > p_k > 0, & \\
n_1, \cdots, n_j, \cdots, n_k, & n_j \in \mathcal{N}, \quad \sum_{j=1}^k n_j = n; \\
r_1, \cdots, r_j, \cdots, r_k, & r_j \in \mathcal{N}, \quad 1 \leqq r_j \leqq \min(n_j, m, p); \\
U_1, \cdots, U_j, \cdots, U_k, & U_j \in \Re^{p\times r_j}, \quad U_j^TU_j = I_{r_j}, \\
\tilde{B}_1, \cdots, \tilde{B}_j, \cdots, \tilde{B}_k, & \tilde{B}_j \in \Re^{r_j\times m} \textit{ positive upper triangular}; \\
\tilde{A}_1, \cdots, \tilde{A}_j, \cdots, \tilde{A}_k, & \tilde{A}_j \in \Re^{n_j\times n_j} \textit{ in } r_j\textit{-balanced form}; \\
D, & D \in \Re^{p\times m}, \quad I - D^TD > 0;
\end{array}$$

in the following way:

If (A, B, C, D) *is partitioned according to* $n_1, \cdots, n_j, \cdots, n_k$, *then*

(i) $B_j = \begin{pmatrix} \tilde{B}_jS^{1/2} \\ 0 \end{pmatrix}$ *where* $S = I - D^TD$, $1 \leqq j \leqq k$;

(ii) $C_j = (R^{1/2} U_j \Delta_j, 0)$ *where* $R = I - DD^T$, $\Delta_j = (\tilde{B}_j \tilde{B}_j^T)^{1/2}$, $1 \leqq j \leqq k$;

(iii) $$A_{jj} = \tilde{A}_j - \frac{1+p_j^2}{p_j}[\operatorname{diag}(\Delta_j^2, 0)]_l - \frac{1+p_j^2}{2p_j}[\operatorname{diag}(\Delta_j^2, 0)]_d - \operatorname{diag}(\tilde{B}_j D^T U_j \Delta_j, 0), \qquad 1 \leqq j \leqq k;$$

(iv) $$A_{ij} = \frac{1}{p_i^2 - p_j^2}(p_j(1-p_i^2)\operatorname{diag}(\tilde{B}_i \tilde{B}_j^T, 0) - p_i(1-p_j^2)\operatorname{diag}(\Delta_i U_i^T U_j \Delta_j, 0)) - \operatorname{diag}(B_i D^T U_j \Delta_j, 0), \qquad 1 \leqq i, j \leqq k, \quad i \neq j.$$

(v) $D \in \Re^{p \times m}$, $\quad I - D^T D > 0$.

Moreover, (A, B, C, D) *as defined in* (2) *is bounded real balanced with bounded real gramian*

$$\Sigma_b = \operatorname{diag}(p_1 I_{n_1}, \cdots, p_j I_{n_j}, \cdots, p_k I_{n_k}).$$

The map Γ_b, *which assigns to each system in* $B_n^{p,m}$ *the realization given in* (2), *is a canonical form.*

Proof. (1) ⇒ (2) Let $(A, B, C, D) \in B_n^{p,m}$ and let $(A_c, B_c, C_c, D_c) \in C_n^{p+m,p+m}$ be the uniquely defined system in Lemma 5.2. If $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ is another system in $B_n^{p,m}$ and $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c) \in C_n^{p+m,p+m}$ is given as in Lemma 5.2, then (A, B, C, D) and $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D})$ are system equivalent if and only if (A_c, B_c, C_c, D_c) and $(\tilde{A}_c, \tilde{B}_c, \tilde{C}_c, \tilde{D}_c)$ are system equivalent. Since (A_c, B_c, C_c, D_c) satisfies the Lyapunov equations (4) and (5) this shows that (A, B, C, D) is bounded real balanced with bounded real gramian Σ_b if and only if (A_c, B_c, C_c, D_c) is balanced with gramian $\Sigma = \Sigma_b = \operatorname{diag}(p_1 I_{n_1}, \cdots, p_j I_{n_j}, \cdots, p_k I_{n_k}) < 1$. We can therefore assume that (A_c, B_c, C_c, D_c) is in the canonical form of Theorem 2.1.

Partitioning the systems in the usual way and using the notation of Theorem 2.1, we obtain

$$B_{c,j} = (B_j S^{-1/2}, p_j C_j^T R^{-1/2}) = \begin{pmatrix} \tilde{B}_{c,j} \\ 0 \end{pmatrix}, \tag{13}$$

with $\tilde{B}_{c,j} \in \Re^{r_j \times (p+m)}$ positive upper triangular and

$$C_{c,j} = \begin{pmatrix} R^{-1/2} C_j \\ p_j S^{-1/2} B_j^T \end{pmatrix} = (U_{c,j} \Delta_{c,j}, 0), \tag{14}$$

where $U_{c,j} \in \Re^{(p+m) \times r_j}$, $U_{c,j}^T U_{c,j} = I_{r_j}$, and $\Delta_{c,j} = (\tilde{B}_{c,j} \tilde{B}_{c,j}^T)^{1/2}$. Since (A_c, B_c, C_c, D_c) is balanced, $B_{c,j} B_{c,j}^T = C_{c,j}^T C_{c,j}$ and therefore we have that

$$B_{c,j} B_{c,j}^T = B_j S^{-1} B_j^T + p_j^2 C_j^T R^{-1} C_j = C_j^T R^{-1} C_j + p_j^2 B_j S^{-1} B_j^T$$

and hence $B_j S^{-1} B_j^T = C_j^T R^{-1} C_j$, which implies that

$$B_{c,j} B_{c,j}^T = (1 + p_j^2) B_j S^{-1} B_j^T. \tag{15}$$

Therefore $r_j = \operatorname{rank}(B_{c,j}) = \operatorname{rank}(\tilde{B}_{c,j}) = \operatorname{rank}(B_j S^{-1/2})$, which implies together with (13) that

$$B_j S^{-1/2} = \begin{pmatrix} \tilde{B}_j \\ 0 \end{pmatrix}$$

with $\tilde{B}_j$ positive upper triangular. This shows (i).

The fact that

$$\operatorname{diag}(\tilde B_j\tilde B_j^T, 0) = B_jS^{-1}B_j^T = C_j^TR^{-1}C_j$$

immediately implies

$$R^{-1/2}C_j = (U_j\Delta_j, 0), \qquad \Delta_j = (\tilde B_j\tilde B_j^T)^{1/2},$$

for a unique $U_j \in \Re^{p\times r_j}$ such that $U_j^TU_j = I_{r_j}$ and hence (ii).

Since by Lemma 5.2

$$A_{c,jj} = \tilde A_{c,j} - \frac{1}{p_j}[B_{c,j}B_{c,j}^T]_l - \frac{1}{2p_j}[B_{c,j}B_{c,j}^T]_d = A_{jj} + B_jS^{-1}D^TC_j,$$

we have using (13), (14), (15)(i), and (15)(ii), and Lemma 5.1,

$$A_{jj} = \tilde A_{c,j} - \frac{1+p_j^2}{p_j}[\operatorname{diag}(\Delta_j^2, 0)]_l - \frac{1+p_j^2}{2p_j}[\operatorname{diag}(\Delta_j^2, 0)]_d - \operatorname{diag}(\tilde B_jD^TU_j\Delta_j, 0).$$

Hence we obtain (iii) by setting $\tilde A_j := \tilde A_{c,j}$.

The parametrization of A_{ij} in (iv) follows immediately from the parametrization of $A_{c,ij}$ in Theorem 2.1 as well as from the expressions for $A_{c,ij}$, $B_{c,j}$, and $C_{c,j}$ in Lemma 5.2.

(2) ⇒ (1) Let (A, B, C, D) be parametrized as in (2). To show $(A, B, C, D) \in C_n^{p,m}$ we construct

$$(A_c, B_c, C_c, D_c) := \left(A + BS^{-1}D^TC, [BS^{-1/2}, PC^TR^{-1/2}], \begin{bmatrix} R^{-1/2}C \\ S^{-1/2}B^TP \end{bmatrix}, D\right),$$

where $P = \operatorname{diag}(p_1I_{n_1}, \cdots, p_jI_{n_j}, \cdots, p_kI_{n_k})$, $S = I - D^TD$, and $R = I - DD^T$.

We must show that (A_c, B_c, C_c, D_c) is in balanced canonical form of Theorem 2.1. To do this we partition the systems in the standard way. Then

$$B_{c,j} = \begin{pmatrix} \tilde B_j & p_j\Delta_jU_j^T \\ 0 & 0 \end{pmatrix},$$

which is positive upper triangular. For $C_{c,j}$ we have

$$C_{c,j} = \begin{pmatrix} U_j\Delta_j & 0 \\ p_j\tilde B_j^T & 0 \end{pmatrix}.$$

Setting

$$\Delta_{c,j}^2 := \tilde B_{c,j}\tilde B_{c,j}^T = \operatorname{diag}((1+p_j^2)\Delta_j^2, 0)$$

and

$$U_{c,j} := \frac{1}{\sqrt{1+p_j^2}}\begin{pmatrix} U_j & 0 \\ p_j\tilde B_j^T\Delta_j^{-1} & 0 \end{pmatrix},$$

we have that

$$C_{c,j} = (U_{c,j}\Delta_{c,j}, 0)$$

with $U_{c,j}^TU_{c,j} = Ir_j$.

That $A_{c,jj}$ and $A_{c,ij}$ are of the required form follows similarly to the derivations in the first part of the proof. Thus (A_c, B_c, C_c, D_c) is parametrized in the balanced canonical form of Theorem 2.1 with gramian P and hence it is in $C_n^{p+m,p+m}$. Since $P^{-1} > P$ Lemma 5.3 implies that $(A, B, C, D) \in C_n^{p,m}$.

To show that (A, B, C, D) is bounded real balanced we must show that P is the minimal and that P^{-1} is the maximal solution to the BRRE. That (A, B, C, D) is bounded real and P is the minimal solution to the BRRE follows since, by construction, P solves the BRRE corresponding to (A, B, C, D) and because of Lemma 5.3 and Proposition 5.1. Repeating the above arguments for the dual system shows that P is also the minimal solution to the dual bounded real Riccati equation. Proposition 5.1 then implies that P^{-1} is the maximal solution to the BRRE. □

Specializing this theorem to the SISO case we obtain the following corollary.

COROLLARY 5.1. *The following two statements are equivalent*:

(1) $b(s) \in TB_n^{1,1}$.

(2) $b(s)$ *has a realization* $(A, b, c, d) \in \Re^{n\times n} \times \Re^{n\times 1} \times \Re^{1\times n} \times \Re^{1\times 1}$ *given by the parameters*:

$$\begin{array}{lll}
1 > p_1 > \cdots > p_j > \cdots > p_k > 0, & & \\
n_1, \cdots, n_j, \cdots, n_k, & n_j \in \mathcal{N}, \quad \sum_{j=1}^{k} n_j = n; & \\
s_1, \cdots, s_j, \cdots, s_k, & s_j = \pm 1, \quad 1 \leqq j \leqq k; & \\
b_1, \alpha(1)_1, \cdots, \alpha(1)_j, \cdots, \alpha(1)_{n_1-1}, & b_1 > 0, \quad \alpha(1)_j > 0, & 1 \leqq j \leqq n_1 - 1; \\
\vdots & & \\
b_i, \alpha(i)_1, \cdots, \alpha(i)_j, \cdots, \alpha(i)_{n_i-1}, & b_i > 0, \quad \alpha(i)_j > 0, & 1 \leqq j \leqq n_i - 1; \\
\vdots & & \\
b_k, \alpha(k)_1, \cdots, \alpha(k)_j, \cdots, \alpha(k)_{n_k-1}, & b_k > 0, \quad \alpha(k)_j > 0, & 1 \leqq j \leqq n_k - 1; \\
d, & d \in \Re, \quad |d| < 1; &
\end{array}$$

in the following way:

(i) $b = (\underbrace{b_1, 0, \cdots, 0}_{n_1}, \cdots, \underbrace{b_j, 0, \cdots, 0}_{n_j}, \cdots, \underbrace{b_k, 0, \cdots, 0}_{n_k})^T$,

(ii) $c = (\underbrace{s_1 b_1, 0, \cdots, 0}_{n_1}, \cdots, \underbrace{s_j b_j, 0, \cdots, 0}_{n_j}, \cdots, \underbrace{s_k b_k, 0, \cdots, 0}_{n_k})$,

(iii) *For* $A =: (A_{ij})_{1 \leqq i,j \leqq k}$ *we have*

(a) *block diagonal entries* A_{jj}, $1 \leqq j \leqq k$:

$$A_{jj} = \begin{pmatrix}
a_{jj} & \alpha(j)_1 & & & & \\
-\alpha(j)_1 & 0 & \alpha(j)_2 & & & \\
 & -\alpha(j)_2 & 0 & \cdot & & 0 \\
 & & \ddots & \ddots & \ddots & \\
 & 0 & & \cdot & 0 & \alpha(j)_{n_j-1} \\
 & & & & -\alpha(j)_{n_j-1} & 0
\end{pmatrix}$$

with $a_{jj} = \dfrac{-b_j^2}{1-d^2}\left(\dfrac{1+p_j^2}{2p_j} + s_j d\right)$;

(b) *off-diagonal blocks* A_{ij}, $1 \leqq i, j \leqq k$, $i \neq j$:

$$A_{ij} = \begin{pmatrix}
a_{ij} & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & & \vdots \\
0 & 0 & \cdots & 0
\end{pmatrix} \quad \textit{with } a_{ij} = \frac{-b_i b_j}{1-d^2}\left(\frac{1 + s_i s_j p_i p_j}{s_i s_j p_i + p_j} + s_j d\right);$$

(iv) $d \in \Re$, $|d| < 1$.

Moreover, (A, b, c, d) *as defined in* (2) *is bounded real balanced with bounded real gramian*

$$\Sigma_b = \operatorname{diag}(p_1 I_{n_1}, \cdots, p_j I_{n_j}, \cdots, p_k I_{n_k}).$$

The map Γ_b, *which assigns to each system in* $B_n^{1,1}$ *the realization given in* (2), *is a canonical form.*

Proof. The corollary follows immediately from the theorem and a straightforward reparametrization of the entries of the b-vector. □

If we analyze the canonical form we have just derived, we can see that it again has the same structure as the canonical forms of the previous sections. The only difference between this canonical form and the previous ones is in the way the parameters enter the A-matrix and that the parameters $p_1, \cdots, p_k$ are bounded by one.

6. Positive real systems. Positive real systems play an important role in many parts of deterministic and stochastic control and system theory. Balanced realizations of positive real systems were introduced in [7] because of their importance in stochastic systems theory (see, e.g., [8], [6]). In this section we will study the question of the parametrization of positive real systems. The approach we take is analogous to the one we have taken in previous sections. We associate a certain type of Riccati equation with positive real systems, the so-called positive real Riccati equations. Balanced realizations for positive real systems are then defined by balancing the minimal solutions of these equations. We could now derive a canonical form for positive real systems in an analogous way to the way the canonical form was derived for asymptotically stable systems in Theorem 2.1. Instead, we use a Moebius transformation to map positive real systems to bounded real systems. In this way we can easily carry to canonical form for bounded real systems over to provide a canonical form and parametrization for positive real systems.

DEFINITION 6.1. A system $(A, B, C, D) \in C_n^{m,m}$ such that $D + D^T > 0$ is called *positive real* if

$$G(iw) + G(-iw)^T > 0, \qquad w \in \Re.$$

We denote by P_n^m the subset of $C_n^{m,m}$ containing all positive real systems. TP_n^m denotes the set of transfer functions of systems in P_n^m.

The derivation of a canonical form for systems in P_n^m will be based on the relationship between positive real and bounded real systems obtained by applying a Moebius transformation to the set $TB_n^{m,m}$ which maps $TB_n^{m,m}$ to TP_n^m (see, e.g., [5]):

$$\begin{aligned} M\colon \quad & TB_n^{m,m} \to TP_n^m \\ & B(s) \;\mapsto P(s) := (I - B(s))^{-1}(I + B(s)). \end{aligned}$$

M is a bijection with inverse

$$\begin{aligned} M^{-1}\colon \quad & TP_n^m \to TB_n^{m,m} \\ & P(s) \mapsto B(s) := (P(s) - I)(P(s) + I)^{-1}. \end{aligned}$$

The corresponding state-space formulae are given in the following proposition.

PROPOSITION 6.1. *The map*

$$\begin{aligned} S_{BP}\colon \quad & B_n^{m,m} \to P_n^m \\ & (A, B, C, D) \mapsto (A + B(I-D)^{-1}C, \sqrt{2}\, B(I-D)^{-1}, \\ & \qquad\qquad \sqrt{2}(I-D)^{-1}C, (I-D)^{-1}(I+D)) \end{aligned}$$

is a bijection with inverse

$$
\begin{aligned}
S_{BP}^{-1}:\quad & P_n^m \to B_n^{m,m}\\
& (A, B, C, D) \mapsto (A - B(I+D)^{-1}C, \sqrt{2}\, B(I+D)^{-1},\\
&\qquad\qquad\qquad\qquad \sqrt{2}\,(I+D)^{-1}C, (D-I)(D+I)^{-1}),
\end{aligned}
$$

such that

- S_{BP} *preserves system equivalence.*
- $P = P^T > 0$ *is a solution to the* BRRE *for* $(A, B, C, D) \in B_n^{m,m}$ *if and only if* $P = P^T > 0$ *is a solution to the positive real Riccati equation* (PRRE)
$$\tilde{A}^T P + P\tilde{A} + (\tilde{C} - \tilde{B}^T P)^T (\tilde{D} + \tilde{D}^T)^{-1} (\tilde{C} - \tilde{B}^T P) = 0$$
for $(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) := S_{BP}((A, B, C, D))$.

This proposition implies that we can define a balanced positive real system analogously to the bounded real case by balancing the minimal solution to the PRRE with the inverse of its maximal solution. Note that this amounts to balancing the minimal solution of the PRRE with the minimal solution of its dual equation.

DEFINITION 6.2. A system $(A, B, C, D) \in P_n^m$ is called *positive real balanced* if

$$P_{\min} = P_{\max}^{-1} = \operatorname{diag}(p_1, \cdots, p_j, \cdots, p_n) =: \Sigma_p,$$

where $P_{\min}$, $P_{\max}$ is the minimal, respectively, maximal solution to the PRRE. Σ_p is called the *positive real gramian* of (A, B, C, D).

As an immediate consequence of the previous proposition, we have that balancing is preserved by the map S_{BP}.

COROLLARY 6.1. *A system* $(A, B, C, D) \in B_n^{m,m}$ *is bounded real balanced with bounded real gramian* Σ_b *if and only if* $S_{BP}((A, B, C, D))$ *is positive real balanced with positive real gramian* $\Sigma_p = \Sigma_b$.

Proposition 6.1 together with the canonical form for bounded real systems in Theorem 5.1 allows us to immediately derive a canonical form for systems in P_n^m.

THEOREM 6.1. *The following two statements are equivalent*:

(1) $P(s) \in TP_n^m$.

(2) $P(s)$ *has a realization* $(A, B, C, D) \in \Re^{n\times n} \times \Re^{n\times m} \times \Re^{m\times n} \times \Re^{m\times m}$ *given by the following parameters*:

$$
\begin{array}{ll}
1 > p_1 > \cdots > p_j > \cdots > p_k > 0, & \\
n_1, \cdots, n_j, \cdots, n_k, & n_j \in \mathcal{N},\quad \sum_{j=1}^k n_j = n;\\
r_1, \cdots, r_j, \cdots, r_k, & r_j \in \mathcal{N},\quad 1 \leqq r_j \leqq \min(n_j, m);\\
U_1, \cdots, U_j, \cdots, U_k, & U_j \in \Re^{m\times r_j}, U_j^T U_j = I_{r_j};\\
\tilde{B}_1, \cdots, \tilde{B}_j, \cdots, \tilde{B}_k, & \tilde{B}_j \in \Re^{r_j\times m}\quad \textit{positive upper triangular};\\
\tilde{A}_1, \cdots, \tilde{A}_j, \cdots, \tilde{A}_k, & \tilde{A}_j \in \Re^{n_j\times n_j}\quad \textit{in } r_j\textit{-balanced form};\\
D_b, & D_b \in \Re^{m\times m},\quad I - D_b^T D_b > 0
\end{array}
$$

in the following way:

If (A, B, C, D) *is partitioned as* $n_1, \cdots, n_j, \cdots, n_k$, *then*,

(i) $B_j = \begin{pmatrix} \sqrt{2}\, \tilde{B}_j S^{1/2} (I - D_b)^{-1} \\ 0 \end{pmatrix}$ *where* $S = I - D_b^T D_b$, $1 \leqq j \leqq k$;

(ii) $C_j = (\sqrt{2}(I - D_b)^{-1} R^{1/2} U_j \Delta_j, 0)$ *where* $R = I - D_b D_b^T$, $\Delta_j = (\tilde{B}_j \tilde{B}_j^T)^{1/2}$, $1 \leqq j \leqq k$;

(iii) $$A_{jj}=\tilde{A}_j-\frac{1+p_j^2}{p_j}[\operatorname{diag}(\Delta_j^2,0)]_l-\frac{1+p_j^2}{2p_j}[\operatorname{diag}(\Delta_j^2,0)]_d$$
$$+\operatorname{diag}(\tilde{B}_jS^{-1/2}(I-D_b^T)(I-D_b)^{-1}R^{1/2}U_j\Delta_j,0)\,,\qquad 1\leqq j\leqq k;$$

(iv) $$A_{ij}=\frac{1}{p_i^2-p_j^2}(p_j(1-p_i^2)\operatorname{diag}(\tilde{B}_i\tilde{B}_j^T,0)-p_i(1-p_j^2)\operatorname{diag}(\Delta_iU_i^TU_j\Delta_j,0))$$
$$+\operatorname{diag}(\tilde{B}_iS^{-1/2}(I-D_b^T)(I-D_b)^{-1}R^{1/2}U_j\Delta_j,0),$$
$$1\leqq i,j\leqq k,\quad i\neq j.$$

(v) $D=(I-D_b)^{-1}(I+D_b)$.

Moreover, (A, B, C, D) as defined (2) is positive real balanced with positive real gramian

$$\Sigma_p=\operatorname{diag}(p_1I_{n_1},\cdots,p_jI_{n_j},\cdots,p_kI_{n_k}).$$

The map Γ_p, which assigns to each system in P_n^m the realization given in (2), is a canonical form.

COROLLARY 6.2. *The following two statements are equivalent:*

(1) $p(s)\in TP_n^1$.

(2) *$p(s)$ has a realization $(A,b,c,d)\in\mathfrak{R}^{n\times n}\times\mathfrak{R}^{n\times 1}\times\mathfrak{R}^{1\times n}\times\mathfrak{R}^{1\times 1}$ given by the parameters:*

$$\begin{array}{lll}
1>p_1>\cdots>p_j>\cdots>p_k>0 & & \\
n_1,\cdots,n_j,\cdots,n_k, & n_j\in\mathcal{N},\quad \sum_{j=1}^k n_j=n; & \\
s_1,\cdots,s_j,\cdots,s_k, & s_j=\pm1,\quad 1\leqq j\leqq k; & \\
b_1,\alpha(1)_1,\cdots,\alpha(1)_j,\cdots,\alpha(1)_{n_1-1}, & b_1>0,\quad \alpha(1)_j>0, & 1\leqq j\leqq n_1-1;\\
\vdots & & \\
b_i,\alpha(i)_1,\cdots,\alpha(i)_j,\cdots,\alpha(i)_{n_i-1}, & b_i>0,\quad \alpha(i)_j>0, & 1\leqq j\leqq n_i-1;\\
\vdots & & \\
b_k,\alpha(k)_1,\cdots,\alpha(k)_j,\cdots,\alpha(k)_{n_k-1}, & b_k>0,\quad \alpha(k)_j>0, & 1\leqq j\leqq n_k-1;\\
d, & d\in\mathfrak{R},\quad d>0; &
\end{array}$$

in the following way:

(i) $b=(\underbrace{b_1,0,\cdots,0}_{n_1},\cdots,\underbrace{b_j,0,\cdots,0}_{n_j},\cdots,\underbrace{b_k,0,\cdots,0}_{n_k})^T,$

(ii) $c=(\underbrace{s_1b_1,0,\cdots,0}_{n_1},\cdots,\underbrace{s_jb_j,0,\cdots,0}_{n_j},\cdots,\underbrace{s_kb_k,0,\cdots,0}_{n_k}),$

(iii) *For $A=:(A_{ij})_{1\leqq i,j\leqq k}$ we have*

(a) *block diagonal entries A_{jj}, $1\leqq j\leqq k$:*

$$A_{jj}=\begin{pmatrix}
a_{jj} & \alpha(j)_1 & & & & \\
-\alpha(j)_1 & 0 & \alpha(j)_2 & & & \\
 & -\alpha(j)_2 & 0 & \cdot & & 0\\
 & & \ddots & \ddots & \ddots & \\
 & 0 & & \cdot & 0 & \alpha(j)_{n_j-1}\\
 & & & & -\alpha(j)_{n_j-1} & 0
\end{pmatrix}$$

with $a_{jj}=\dfrac{-b_j^2}{4dp_j}(1-s_jp_j)^2;$

(b) *off-diagonal blocks* A_{ij}, $1 \leqq i, j \leqq k$, $i \neq j$:

$$A_{ij} = \begin{pmatrix} a_{ij} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \quad \textit{with } a_{ij} = \frac{-b_i b_j}{2d(s_i s_j p_i + p_j)}(1 - s_i p_i)(1 - s_j p_j);$$

(iv) $d \in \Re$, $d > 0$.

Moreover, (A, b, c, d) *as defined in* (2) *is positive real balanced with positive real gramian*

$$\Sigma_p = \operatorname{diag}(p_1 I_{n_1}, \cdots, p_j I_{n_j}, \cdots, p_k I_{n_k}).$$

The map Γ_p, *which assigns to each system in* P_n^1 *the realization given in* (2), *is a canonical form.*

Note that in the derivation of the corollary from the theorem we have used an obvious reparametrization of the b-vector to obtain the usual statement of the result. The previous theorem and corollary show the by-now-expected structure of the canonical form.

7. Minimum-phase systems. The last class of systems for which we would like to derive a canonical form is the class of minimum-phase systems. A minimum-phase system is an asymptotically stable system whose inverse system is also asymptotically stable. Minimum-phase systems, therefore, are precisely those systems whose transfer functions are real rational functions in H^∞, which are units, i.e., invertible in H^∞. These systems are of importance in many different areas. In this section, however, we are mainly interested in results concerning minimum-phase systems that are motivated by problems in stochastic system theory, since those results allow us to use the canonical form derived for positive real systems to obtain a canonical form for minimum-phase systems.

DEFINITION 7.1. A system $(A, B, C, D) \in C_n^{m,m}$ such that D is invertible is called *minimum phase* if $A - BD^{-1}C$ has its eigenvalues in the open left halfplane. We denote by M_n^m the subset of $C_n^{m,m}$ containing all minimum-phase systems. TM_n^m denotes the set of transfer functions of systems in M_n^m.

The role minimum-phase systems play in the spectral factorization problem is indicated in the following proposition, which summarizes some standard results (see, e.g., [8], [32]).

PROPOSITION 7.1.

(1) *Let* $(A, B, C, D) \in C_n^{m,m}$ *such that* $D + D^T > 0$ *and assume that there exists a solution* $P = P^T > 0$ *to the corresponding* PRRE. *Then the following statements are equivalent*:

(i) $(A, B, C, D) \in P_n^m$.

(ii) *The minimal solution* $P_{\min}$ *and the maximal solution* $P_{\max}$ *to the* PRRE *are such that* $0 < P_{\min} < P_{\max}$.

(iii) *There exists a solution* P_0 *to the* PRRE *such that the eigenvalues of*

$$A - B(D + D^T)^{-1}(C - B^T P_0)$$

are in the open left halfplane.

If (iii) *is satisfied, then* $P_0 = P_{\min}$.

(2) *Let* $P(s) \in TP_n^m$ *with realization* $(A, B, C, D) \in P_n^m$. *Then*

$$P(iw) + P(-iw)^T = M(-iw)^T M(iw), \qquad w \in \Re,$$

for some $M(s) \in TM_n^m$. A minimal realization of $M(s)$ is given by

$$(A, B, C, D_m^{-T}(C - B^T P_{\min}), D_m),$$

where $P_{\min}$ is the minimal solution to the PRRE *corresponding to (A, B, C, D) and $D_m \in \mathfrak{R}^{m\times m}$ is such that $D_m^T D_m = D + D^T$.*

This proposition suggests how we can relate positive real systems to minimum-phase systems. The precise relationship is given in the following proposition where for a minimum-phase system with a given D-matrix we uniquely define an associated positive real system.

PROPOSITION 7.2. *For $D \in \mathfrak{R}^{m\times m}$, invertible, let*

$$M_{n,D}^m = \{(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) \in M_n^m \mid \tilde{D} = D\},$$
$$P_{n,D}^m = \{(\tilde{A}, \tilde{B}, \tilde{C}, \tilde{D}) \in P_n^m \mid \tilde{D} = \tfrac{1}{2} D^T D\},$$

with $TM_{n,D}^m$ and $TP_{n,D}^m$ the sets of corresponding transfer functions.

The map

$$S_{PM,D}\colon \quad P_{n,D}^m \to M_{n,D}^m$$
$$(A_p, B_p, C_p, D_p) \mapsto (A_m, B_m, C_m, D_m) := (A_p, B_p, D^{-T}(C_p - B_p^T P_{\min}), D),$$

where $P_{\min}$ is the minimal solution to the PRRE *corresponding to (A_p, B_p, C_p, D_p), is a bijection. Its inverse is given by*

$$S_{PM,D}^{-1}\colon \quad M_{n,D}^m \to P_{n,D}^m$$
$$(A_m, B_m, C_m, D_m) \mapsto (A_p, B_p, C_p, D_p) := (A_m, B_m, D^T C_m + B_m^T P, \tfrac{1}{2} D^T D),$$

where P is the solution to the Lyapunov equation

$$A_m^T P + P A_m = -C_m^T C_m. \tag{16}$$

Moreover, $S_{PM,D}$ preserves system equivalence.

Proof. By Proposition 7.1 the map $S_{PM,D}$ is well defined. To show that $S_{PM,D}^{-1}$ is well defined we first must show that for $(A_m, B_m, C_m, D_m) \in M_{n,D}^m$ the system $(A_p, B_p, C_p, D_p) = S_{PM,D}^{-1}((A_p, B_p, C_p, D_p))$ is in $C_n^{m,m}$.

First note that P solves the PRRE corresponding to (A_p, B_p, C_p, D_p). Let $P(s) = C_p(sI - A_p)^{-1}B_p + D_p$ and $M(s) = C_m(sI - A_m)^{-1}B_m + D_m$; then standard algebraic manipulations show that

$$P(iw) + P(-iw)^T = M(-iw)^T M(iw), \qquad w \in \mathfrak{R}.$$

Since by assumption $M(s)$ has McMillan degree n this implies that $P(s)$ also has McMillan degree n. This shows that $(A_p, B_p, C_p, D_p) \in C_n^{m,m}$.

$(A_p, B_p, C_p, D_p) \in P_n^m$ follows from Proposition 7.1 as

$$\tilde{A} := A_p - B_p(D_p + D_p^T)^{-1}(C_p - B_p^T P) = A_m - B_m D_m^{-1} C_m$$

has all its eigenvalues in the open left halfplane. It is now straightforward to verify that $S_{PM,D}$ is in fact a bijection and preserves system equivalence. □

We now define a minimum-phase system to be minimum-phase balanced if its corresponding positive real system is positive real balanced.

DEFINITION 7.2. A system $(A, B, C, D) \in M_n^m$ is called *minimum-phase balanced* if the system $S_{PM,D}^{-1}((A, B, C, D))$ is positive real balanced. The *minimum-phase gramian* Σ_m of (A, B, C, D) is defined to be the positive real gramian Σ_p of $S_{PM,D}^{-1}((A, B, C, D))$, i.e., $\Sigma_m = \Sigma_p$.

A canonical form for minimum-phase systems can now be derived using Proposition 7.2 to carry the canonical form for positive real systems over to the class of minimum-phase systems.

THEOREM 7.1. *The following two statements are equivalent*:

(1) $M(s) \in TM_n^m$.

(2) $M(s)$ *has a realization* $(A, B, C, D) \in \mathfrak{R}^{n\times n} \times \mathfrak{R}^{n\times m} \times \mathfrak{R}^{m\times n} \times \mathfrak{R}^{m\times m}$ *given by the parameters*:

$$\begin{array}{ll}
1 > p_1 > \cdots > p_j > \cdots > p_k > 0, & \\
n_1, \cdots, n_j, \cdots, n_k, & n_j \in \mathcal{N}, \quad \sum_{j=1}^k n_j = n; \\
r_1, \cdots, r_j, \cdots, r_k, & r_j \in \mathcal{N}, \quad 1 \leqq r_j \leqq \min(n_j, m); \\
U_1, \cdots, U_j, \cdots, U_k, & U_j \in \mathfrak{R}^{m\times r_j}, \quad U_j^T U_j = I_{r_j}; \\
\tilde{B}_1, \cdots, \tilde{B}_j, \cdots, \tilde{B}_k, & \tilde{B}_j \in \mathfrak{R}^{r_j \times m} \quad \textit{positive upper triangular}; \\
\tilde{A}_1, \cdots, \tilde{A}_j, \cdots, \tilde{A}_k, & \tilde{A}_j \in \mathfrak{R}^{n_j\times n_j} \quad \textit{in } r_j\textit{-balanced form}; \\
D, & D \in \mathfrak{R}^{m\times m}, \quad D \textit{ invertible}
\end{array}$$

in the following way:

If (A, B, C, D) *is partitioned according to* $n_1, \cdots, n_j, \cdots, n_k$, *then*,

(i) $B_j = \begin{pmatrix} \tilde{B}_j (D^T D)^{1/2} \\ 0 \end{pmatrix}, \quad 1 \leqq j \leqq k;$

(ii) $C_j = D^{-T}(D^T D)^{1/2}(U_j \Delta_j - p_j \tilde{B}_j^T, 0)$ *where* $\Delta_j = (\tilde{B}_j \tilde{B}_j^T)^{1/2}, \quad 1 \leqq j \leqq k;$

(iii) $$A_{jj} = \tilde{A}_j - \frac{1+p_j^2}{p_j}[\operatorname{diag}(\Delta_j^2, 0)]_l - \frac{1+p_j^2}{2p_j}[\operatorname{diag}(\Delta_j^2, 0)]_d + \operatorname{diag}(\tilde{B}_j U_j \Delta_j, 0), \quad 1 \leqq j \leqq k;$$

(iv) $$A_{ij} = \frac{1}{p_i^2 - p_j^2}(p_j(1-p_i^2)\operatorname{diag}(\tilde{B}_i \tilde{B}_j^T, 0) - p_i(1-p_j^2)\operatorname{diag}(\Delta_i U_i^T U_j \Delta_j, 0)) + \operatorname{diag}(\tilde{B}_i U_j \Delta_j, 0), \quad 1 \leqq i, j \leqq k, \quad i \neq j.$$

(v) $D \in \mathfrak{R}^{m\times m}$, *invertible.*

Moreover, (A, B, C, D) *as defined in* (2) *is minimum-phase balanced with minimum-phase gramian*

$$\Sigma_m = \operatorname{diag}(p_1 I_{n_1}, \cdots, p_j I_{n_j}, \cdots, p_k I_{n_k}).$$

The map Γ_m, *which assigns to each system in* M_n^m *the realization given in* (2), *is a canonical form.*

COROLLARY 7.1. *The following two statements are equivalent*:

(1) $m(s) \in TM_n^1$.

(2) $m(s)$ *has a realization* $(A, b, c, d) \in \mathfrak{R}^{n\times n} \times \mathfrak{R}^{n\times 1} \times \mathfrak{R}^{1\times n} \times \mathfrak{R}^{1\times 1}$ *given by the parameters*:

$$\begin{array}{ll}
1 > p_1 > \cdots > p_j > \cdots > p_k > 0, & \\
n_1, \cdots, n_j, \cdots, n_k, & n_j \in \mathcal{N}, \quad \sum_{j=1}^k n_j = n; \\
s_1, \cdots, s_j, \cdots, s_k, & s_j = \pm 1, \quad 1 \leqq j \leqq k; \\
b_1, \alpha(1)_1, \cdots, \alpha(1)_j, \cdots, \alpha(1)_{n_1-1}, & b_1 > 0, \quad \alpha(1)_j > 0, \quad 1 \leqq j \leqq n_1 - 1; \\
\vdots & \\
b_i, \alpha(i)_1, \cdots, \alpha(i)_j, \cdots, \alpha(i)_{n_i-1}, & b_i > 0, \quad \alpha(i)_j > 0, \quad 1 \leqq j \leqq n_i - 1; \\
\vdots & \\
b_k, \alpha(k)_1, \cdots, \alpha(k)_j, \cdots, \alpha(k)_{n_k-1}, & b_k > 0, \quad \alpha(k)_j > 0, \quad 1 \leqq j \leqq n_k - 1; \\
d, & d \in \mathfrak{R}, \quad d \neq 0
\end{array}$$

in the following way:

(i) $b=(\underbrace{b_1,0,\cdots,0}_{n_1},\cdots,\underbrace{b_j,0,\cdots,0}_{n_j},\cdots,\underbrace{b_k,0,\cdots,0}_{n_k})^T$,

(ii) $c=\frac{1}{d}((\underbrace{(s_1-p_1)b_1,0,\cdots,0}_{n_1},\cdots,\underbrace{(s_j-p_j)b_j,0,\cdots,0}_{n_j},\cdots,$

$\underbrace{(s_k-p_k)b_k,0,\cdots,0}_{n_k})$,

(iii) *For* $A=:(A_{ij})_{1\leqq i,j\leqq k}$ *we have*

(a) *block diagonal entries* A_{jj}, $1\leqq j\leqq k$:

$$A_{jj}=\begin{pmatrix} a_{jj} & \alpha(j)_1 & & & & \\ -\alpha(j)_1 & 0 & \alpha(j)_2 & & & \\ & -\alpha(j)_2 & 0 & \cdot & & 0 \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \cdot & 0 & \alpha(j)_{n_j-1} \\ & & & & -\alpha(j)_{n_j-1} & 0 \end{pmatrix}$$

with $a_{jj}=\dfrac{-b_j^2}{2d^2p_j}(1-s_jp_j)^2$;

(b) *off-diagonal blocks* A_{ij}, $1\leqq i,j\leqq k$, $i\neq j$:

$$A_{ij}=\begin{pmatrix} a_{ij} & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix} \quad \textit{with } a_{ij}=\frac{-b_ib_j}{d^2(s_is_jp_i+p_j)}(1-s_ip_i)(1-s_jp_j);$$

(iv) $d\in\Re$, $d\neq 0$.

Moreover, (A, b, c, d) *as defined in* (2) *is minimum-phase balanced with minimum-phase gramian*

$$\Sigma_m=\operatorname{diag}(p_1I_{n_1},\cdots,p_jI_{n_j},\cdots,p_kI_{n_k}).$$

The map Γ_m, *which assigns to each system in* M_n^1 *the realization given in* (2), *is a canonical form.*

While the canonical form we have just derived has a structure that is very similar to the canonical forms derived for the other classes of systems, there is a certain loss of symmetry in this canonical form. This is most apparent in the case of single-input single-output systems. In the previously derived canonical forms, the entries of the b-vector and c-vector had the same modulus. Moreover, the canonical forms were sign-symmetric (see § 10 for a precise definition). The sign-symmetry is lost in the above canonical form for single-input single-output minimum-phase systems. By standard arguments it can be shown that all single-input single-output minimum-phase systems have a sign-symmetric realization; the canonical form for single-input single-output asymptotically stable systems of Corollary 2.1 restricted to minimum-phase systems is such an example. It is, however, not clear whether it is possible to find a "balancing scheme" for minimum-phase systems such that the corresponding canonical form is sign-symmetric for single-input single-output systems and at the same time leads to a parametrization whose parameter space has the desirable geometric structure of the canonical form presented here.

8. Discrete-time systems. The canonical forms derived in the previous sections dealt with continuous-time systems. In this section we will show that the standard technique of bilinearly transforming continuous-time systems to discrete-time systems can be applied to derive canonical forms for various classes of discrete-time systems. We first define the classes of systems that will be considered.

DEFINITION 8.1. Let $(A, B, C, D) \in \mathfrak{R}^{n\times n} \times \mathfrak{R}^{n\times m} \times \mathfrak{R}^{p\times n} \times \mathfrak{R}^{p\times m}$ and $G(z) = C(zI-A)^{-1}B+D$.

(1) If all eigenvalues of A are in the open unit disk, then (A, B, C, D) is called *discrete-time asymptotically stable.* The set of discrete-time asymptotically stable systems in $L_n^{p,m}$ is denoted by $D_n^{p,m}$ with $TD_n^{p,m}$ the corresponding set of transfer functions.

(2) A system $(A, B, C, D) \in D_n^{p,m}$ is called *discrete-time bounded real* if

$$I - G(e^{-i\theta})^T G(e^{i\theta}) > 0, \qquad \theta \in [0, 2\pi].$$

The set of discrete-time bounded real systems in $D_n^{m,m}$ is denoted by $DB_n^{p,m}$ with $TDB_n^{p,m}$ the corresponding set of transfer functions.

(3) A system $(A, B, C, D) \in D_n^{m,m}$ is called *discrete-time positive real* if

$$G(e^{-i\theta})^T + G(e^{i\theta}) > 0, \qquad \theta \in [0, 2\pi].$$

The set of discrete-time positive real systems in $D_n^{p,m}$ is denoted by DP_n^m with TDP_n^m the corresponding set of transfer functions.

(4) A system $(A, B, C, D) \in D_n^{m,m}$ is called *discrete-time minimum phase* if

$$\tilde{G}(z) := G(z)^{-1} \in TD_n^{m,m}.$$

The set of discrete-time minimum-phase systems in $D_n^{m,m}$ is denoted by DM_n^m with TDM_n^m the corresponding set of transfer functions.

(5) A system $(A, B, C, D) \in D_n^{m,m}$ is called *discrete-time allpass* if for some $\sigma > 0$,

$$G(e^{i\theta}) G(e^{-i\theta})^T = \sigma^2 I, \qquad \theta \in [0, 2\pi].$$

The set of discrete-time allpass systems in $D_n^{m,m}$ is denoted by DA_n^m with TDA_n^m the corresponding set of transfer functions.

The following proposition summarizes some basic results on the bilinear transformation.

PROPOSITION 8.1. *The transformation*

$$TU_n^{p,m}: TC_n^{p,m} \to TD_n^{p,m},$$

$$G_c(s) \mapsto G_d(z) := G_c\left(\frac{z-1}{z+1}\right)$$

is a bijection with inverse

$$(TU_n^{p,m})^{-1}: TD_n^{p,m} \to TC_n^{p,m},$$

$$G_d(z) \mapsto G_c(s) := G_d\left(\frac{1+s}{1-s}\right),$$

which induces a bijection between $TB_n^{p,m}$ and $TDB_n^{p,m}$. If $p = m$ then $TU_n^{m,m}$ induces a bijection between TA_n^m and TDA_n^m, TP_n^m and TDP_n^m, as well as TM_n^m and TDM_n^m.

This mapping also has a formulation in terms of state-space systems, which is given in the next proposition [3], [10], [22].

PROPOSITION 8.2. *The transformation*

$$SU_n^{p,m}: \quad C_n^{p,m} \to D_n^{p,m},$$
$$(A_c, B_c, C_c, D_c) \mapsto (A_d, B_d, C_d, D_d),$$
$$(A_d, B_d, C_d, D_d) := ((I-A_c)^{-1}(I+A_c), \sqrt{2}(I-A_c)^{-1}B_c, \sqrt{2}\, C_c(I-A_c)^{-1}, D_c + C_c(I-A_c)^{-1}B_c)$$

is a bijection with inverse

$$(SU_n^{p,m})^{-1}: \quad D_n^{p,m} \to C_n^{p,m},$$
$$(A_d, B_d, C_d, D_d) \mapsto (A_c, B_c, C_c, D_c),$$
$$(A_c, B_c, C_c, D_c) := ((I+A_d)^{-1}(A_d - I), \sqrt{2}(I+A_d)^{-1}B_d, \sqrt{2}\, C_d(I+A_d)^{-1}, D_d - C_d(I+A_d)^{-1}B_d),$$

which induces a bijection between $B_n^{p,m}$ and $DB_n^{p,m}$. If $p = m$, then $SU_n^{m,m}$ induces a bijection between A_n^m and DA_n^m, P_n^m and DP_n^m, as well as M_n^m and DM_n^m. The map $SU_n^{p,m}$ preserves system equivalence as well as sign-symmetry of state-space realizations if $p = m$; i.e., for $(A_c, B_c, C_c, D_c) = (SU_n^{m,m})^{-1}((A_d, B_d, C_d, D_d))$, $(A_d, B_d, C_d, D_d) \in D_n^{m,m}$ we have

$$A_c = SA_c^T S, \qquad B_c = SC_c^T$$

if and only if

$$A_d = SA_d^T S, \qquad B_d = SC_d^T,$$

for some $S = \operatorname{diag}(\pm 1, \cdots, \pm 1)$.

The previous proposition allows us to carry over the canonical forms for continuous-time asymptotically stable systems to the discrete-time case.

THEOREM 8.1. *If Γ, Γ_a, Γ_b, Γ_p, and Γ_m are the canonical forms for the sets $C_n^{p,m}$, A_n^m, $B_n^{p,m}$, P_n^m, and M_n^m as defined in the previous sections, then $D\Gamma := SU_n^{p,m}\Gamma(SU_n^{p,m})^{-1}$, $D\Gamma_a := SU_n^{m,m}\Gamma_a(SU_n^{m,m})^{-1}$, $D\Gamma_b := SU_n^{p,m}\Gamma_b(SU_n^{p,m})^{-1}$, $D\Gamma_p := SU_n^{m,m}\Gamma_p(SU_n^{m,m})^{-1}$, and $D\Gamma_m := SU_n^{m,m}\Gamma_m(SU_n^{m,m})^{-1}$ are canonical forms for the sets $D_n^{p,m}$, DA_n^m, $DB_n^{p,m}$, DP_n^m, and DM_n^m.*

Remark 8.1. Analogously to the continuous-time case we can introduce balancing techniques for the various classes of discrete-time systems by balancing solutions to the corresponding discrete-time Lyapunov and Riccati equations. Since the map $SU_n^{p,m}$ leaves such solutions invariant (see, e.g., [3], [10], [22]), $SU_n^{p,m}$, in fact, preserves balancing. Hence the canonical forms for discrete-time systems introduced in Theorem 8.1 are therefore in terms of balanced representations.

9. Model reduction. One of the main advantages of balanced representations is that they can be used for a very straightforward method of model reduction. Moore [19] has, in fact, introduced balanced realizations for stable linear systems to have an efficient way of performing model reduction. His scheme was based on the following state-space projection method. Consider an n-dimensional balanced system (A, B, C, D) and partition it conformally as

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \quad C = (C_1, C_2),$$

such that for $1 \leqq N < n$, $A_{11} \in \mathfrak{R}^{N \times N}$, $B_1 \in \mathfrak{R}^{N \times m}$, and $C_1 \in \mathfrak{R}^{p \times N}$. The principal subsystem (A_{11}, B_1, C_1, D) is then considered to be an approximant of (A, B, C, D).

Pernebo and Silverman [29] were the first to show that the approximant of a balanced system in $C_n^{p,m}$ is again balanced, minimal, and asymptotically stable. Their result, however, assumes that truncation does not occur between states corresponding to repeated singular values. Otherwise the approximant may no longer be asymptotically stable and minimal. Analogous results with the same restriction were shown in [7] for positive real systems, in [16] for Riccati balanced systems, and in [28] for bounded real systems.

We now suggest a model reduction technique that is more suitable to our particular setup than the method discussed above. Rather than working with the state-space systems directly, we perform the model reduction by reducing the parameters corresponding to a certain system.

We have seen that associated with each system is a set of parameters as follows:

$$\begin{array}{ll}
\sigma_1 > \cdots > \sigma_j > \cdots > \sigma_k > 0, & \\
n_1, \cdots, n_j, \cdots, n_k, & n_j \in \mathcal{N}, \quad \sum_{j=1}^{k} n_j = n; \\
r_1, \cdots, r_j, \cdots, r_k, & r_j \in \mathcal{N}, \quad 1 \leqq r_j \leqq \min(n_j, m, p); \\
U_1, \cdots, U_j, \cdots, U_k, & U_j \in \mathfrak{R}^{p \times r_j}, \quad U_j^T U_j = I_{r_j}; \\
\tilde{B}_1, \cdots, \tilde{B}_j, \cdots, \tilde{B}_k, & \tilde{B}_j \in \mathfrak{R}^{r_j \times m} \quad \text{positive upper triangular}; \\
\tilde{A}_1, \cdots, \tilde{A}_j, \cdots, \tilde{A}_k, & \tilde{A}_j \in \mathfrak{R}^{n_j \times n_j} \quad \text{in } r_j\text{-balanced form}; \\
D, & D \in \mathfrak{R}^{p \times m}
\end{array}$$

A reduced-order system of degree N can now easily be defined by retaining of these parameters only those that correspond to the first N states; i.e., if j_N is such that $n_1 + \cdots + n_{j_N} < N \leqq n_1 + \cdots + n_{jn} + h_{j_N+1}$, then take the following parameters:

$$\begin{array}{ll}
\sigma_1, \cdots, \sigma_{j_N}, \sigma_{j_N+1}, & \\
n_1, \cdots, n_{j_N}, Pn_{j_N+1}, & Pn_{j_N+1} := N - (n_1 + \cdots + n_{j_N}); \\
U_1, \cdots, U_{j_N}, PU_{j_N+1}, PU_{j_N+1}, & \text{the first } \min(r_{j_N+1}, Pn_{j_N+1}) \\
 & \text{columns of } U_{j_N+1}. \\
\tilde{B}_1, \cdots, \tilde{B}_{j_N}, P\tilde{B}_{j_N+1}, & P\tilde{B}_{j_N+1} \text{ the first } \min(r_{j_N+1}, Pn_{j_N+1}) \\
 & \text{rows of } \tilde{B}_{j_N+1}; \\
\tilde{A}_1, \cdots, \tilde{A}_{j_N}, P\tilde{A}_{j_N+1}, P\tilde{A}_{j_N+1}, & \text{the principal submatrix of} \\
 & \tilde{A}_{j_N+1} \text{ of size } Pn_{j_N+1}.
\end{array}$$

By the parametrization results of the previous sections these parameters define a unique reduced-order system that is in the same class of the systems as the original system. We can call this scheme a *parameter projection method.*

We can summarize this method in the following theorem.

THEOREM 9.1. *If* (A, B, C, D) *is a continuous-time system in* $C_n^{p,m}(A_n^m, L_n^{p,m}, B_n^{p,m}, P_n^m, M_n^m)$ *given in the canonical form of Theorem* 2.1 (3.1, 4.1, 5.1, 6.1, 7.1) *with parameters*

$$\begin{array}{ll}
\sigma_1 > \cdots > \sigma_j > \cdots > \sigma_k > 0, & \\
n_1, \cdots, n_j, \cdots, n_k, & n_j \in \mathcal{N}, \quad \sum_{j=1}^{k} n_j = n; \\
r_1, \cdots, r_j, \cdots, r_k, & r_j \in \mathcal{N}, \quad 1 \leqq r_j \leqq \min(n_j, m, p); \\
U_1, \cdots, U_j, \cdots, U_k, & U_j \in \mathfrak{R}^{p \times r_j}, \quad U_j^T U_j = I_{r_j}; \\
\tilde{B}_1, \cdots, \tilde{B}_j, \cdots, \tilde{B}_k, & \tilde{B}_j \in \mathfrak{R}^{r_j \times m} \quad \textit{positive upper triangular}; \\
\tilde{A}_1, \cdots, \tilde{A}_j, \cdots, \tilde{A}_k, & \tilde{A}_j \in \mathfrak{R}^{n_j \times n_j} \quad \textit{in } r_j\textit{-balanced form}; \\
D, & D \in \mathfrak{R}^{p \times m};
\end{array}$$

then each N-dimensional system, $1 \leqq N < n$, *obtained by the balanced parameter space model reduction scheme, i.e., each system given by the parameters*

$$\begin{array}{ll}
\sigma_1, \cdots, \sigma_{j_N}, \sigma_{j_N+1}, & \\
n_1, \cdots, n_{j_N}, Pn_{j_N+1}, & Pn_{j_N+1} := N - (n_1 + \cdots + n_{j_N}); \\
U_1, \cdots, U_{j_N}, PU_{j_N+1}, & PU_{j_N+1} \textit{ the first } \min(r_{j_N+1}, Pn_{j_N+1}) \\
 & \textit{columns of } U_{j_N+1}; \\
\tilde{B}_1, \cdots, \tilde{B}_{j_N}, P\tilde{B}_{j_N+1}, & P\tilde{B}_{j_N+1} \textit{ the first } \min(r_{j_N+1}, Pn_{j_N+1}) \\
 & \textit{rows of } \tilde{B}_{j_N+1}; \\
\tilde{A}_1, \cdots, \tilde{A}_{j_N}, P\tilde{A}_{j_N+1}, & P\tilde{A}_{j_N+1} \textit{ the principal submatrix of} \\
D, & \tilde{A}_{j_N+1} \textit{ of size } Pn_{j_N+1},
\end{array}$$

and parametrized as in Theorem 2.1 (3.1, 4.1, 5.1, 6.1, 7.1) *is again in canonical form and therefore in the same class of systems, i.e., in* $C_N^{p,m}$ $(A_N^m, L_N^{p,m}, B_n^{p,m}, P_N^m, M_N^m)$.

From this result we can easily obtain a result concerning model reduction using the state-space projection method.

COROLLARY 9.1. *If* (A, B, C, D) *is a continuous-time system in* $C_n^{p,m}$ $(A_n^m, L_n^{p,m}, B_n^{p,m}, P_n^m, M_n^m)$ *given in the canonical form of Theorem* 2.1 (3.1, 4.1, 5.1, 6.1, 7.1) *then each N-dimensional principal subsystem of* (A, B, C, D) *is again in canonical form and therefore in the same class of systems, i.e., in* $C_N^{p,m}(A_N^m, L_N^{p,m}, B_n^{p,m}, P_N^m, M_N^m)$ *if for some* $0 \leqq j_N \leqq k-1$,

$$n_1 + \cdots + n_{j_N} + r_{j_N+1} \leqq N \leqq n_1 + \cdots + n_{j_N} + n_{j_N+1},$$

where we set $n_0 = 0$.

Proofs. The result follows by inspection since the reduced-order system is parametrized by a set of parameters as in Theorem 9.1. □

Because of the particular nature of the nonuniqueness of balanced realizations Corollary 9.1 is, in fact, more general than the results in [29], [7], [16], and [28], where it is assumed that truncation occurs at a point of nonrepeated singular values. We can recover these results immediately in the following corollary.

COROLLARY 9.2. *If* (A, B, C, D) *is a continuous-time system given in* $C_n^{p,m}(L_n^{p,m}, B_n^{p,m}, P_n^m, M_n^m)$ *that is Lyapunov balanced (Riccati balanced, bounded real balanced, positive real balanced, minimum-phase balanced) with gramian* $\Sigma = \operatorname{diag}(\sigma_1 I_{n_1}, \cdots, \sigma_k I_{n_k})$, *then the N-dimensional principal subsystem is in the same class of systems, i.e.,* $C_N^{p,m}(L_N^{p,m}, B_n^{p,m}, P_N^m, M_N^m)$, *if* $N = n_1 + n_2 + \cdots + n_{j_0}$, *for some* $j_0 = 1, \cdots, k$.

We obtain a very general model reduction result if either the input or the output dimension of the system is one. Then we have that the r_j parameters are also one, and hence the condition Corollary 9.1 is always satisfied.

COROLLARY 9.3. *Assume that* $\min(p, m) = 1$. *If* (A, B, C, D) *is a continuous-time system in* $C_n^{p,m}$ $(A_n^m, L_n^{p,m}, B_n^{p,m}, P_n^m, M_n^m)$ *given in the canonical form of Theorem* 2.1 (3.1, 4.1, 5.1, 6.1, 7.1), *then each N-dimensional principal subsystem of* (A, B, C, D) *is again in canonical form and therefore in the same class of systems, i.e., in* $C_N^{p,m}$ $(A_N^m, L_N^{p,m}, B_n^{p,m}, P_N^m, M_N^m)$.

Remark 9.1. Note that for the particular type of canonical forms presented in this paper we cannot, in general, expect to have a model reduction result based on the state-space projection method by which truncation can occur at an arbitrary place. This is due to the fact that we use $(\tilde{B}_j\tilde{B}_j^T)^{1/2}$ as a parameter in the C-matrix and that $(\tilde{B}_j\tilde{B}_j^T)^{1/2}$ has no specific structure. In the case of the canonical forms presented in [22] and [27], $(\tilde{B}_j\tilde{B}_j^T)^{1/2}$ was constrained to be diagonal and we have the general model reduction property.

In [1] and [22] a model reduction technique was suggested for balanced discrete-time systems in $D_n^{p,m}$ by carrying a discrete-time system over to a continuous-time system using the map $SU_n^{p,m}$. This corresponding continuous-time system is reduced to a lower-order system which is then mapped back to a discrete-time system using the inverse mapping. Thereby we obtain a lower-order approximant to the discrete-time system. The following corollary shows how it is possible to obtain in this way a discrete-time version of Corollary 9.1.

COROLLARY 9.4. *Let (A_d, B_d, C_d, D_d) be in one of the following classes of discrete-time asymptotically stable systems: $D_n^{p,m}$, DA_n^m, $DB_n^{p,m}$, DP_n^m, and DM_n^m. Assume that (A_d, B_d, C_d, D_d) is given in the corresponding canonical form. Let $(\hat{A}_c, \hat{B}_c, \hat{C}_c, \hat{D}_c)$ be the N-dimensional principal subsystem of $(A_c, B_c, C_c, D_c) = (SU_n^{p,m})^{-1}((A_d, B_d, C_d, D_d))$; then $(\hat{A}_d, \hat{B}_d, \hat{C}_d, \hat{D}_d) := SU_n^{p,m}((\hat{A}_c, \hat{B}_c, \hat{C}_c, \hat{D}_c))$ is in the corresponding subclass of N-dimensional systems, if*

$$n_1 + \cdots + n_{j_N} + r_{j_N+1} \leqq N \leqq n_1 + \cdots + n_{j_N} + n_{j_N+1}.$$

Clearly, all the other continuous-time results in this section can be carried over to discrete-time systems in the same way.

10. Final remarks. The paper dealt with canonical forms and parametrizations for the following classes of linear systems of fixed dimensions: asymptotically stable systems, allpass systems, the general class of minimal systems, positive real systems, bounded real systems, and minimum-phase systems. All the canonical forms are given in terms of balanced realizations for the particular class of systems. Several aspects of these parametrizations were discussed including model reduction.

It was pointed out that all the canonical forms have a similar structure. Only the way the parameters enter the entries of the system matrices determines whether or not a system belongs to a certain class of systems. We are going to make a few further remarks concerning common properties of the various canonical forms.

Sign-symmetry and Cauchy index. With the exception of minimum-phase systems all scalar systems that are given in one of the previously derived canonical forms have the so-called sign-symmetry property, i.e.,

$$A^T = SAS, \qquad b = Sc^T,$$

where S is a diagonal matrix whose diagonal terms are ±1. In particular, if

$$n_1, \cdots, n_j, \cdots, n_k, \qquad s_1, \cdots, s_j, \cdots, s_k$$

are the usual structural parameters of a scalar system given in one of the canonical forms of Corollaries 2.1, 3.1, 4.1, 5.1, 6.2, 7.1, then the sign-symmetry matrix S is

$$S = \operatorname{diag}(s_1\hat{I}_{n_1}, \cdots, s_j\hat{I}_{n_j}, \cdots, s_k\hat{I}_{n_k}),$$

where $\hat{I}_{n_j} = \operatorname{diag}(+1, -1, +1, \cdots, (-1)^{n_j+1}) \in \Re^{n_j \times n_j}$. Note that by Proposition 8.2 the canonical form of a discrete-time system is sign symmetric if its corresponding continuous-time system is sign symmetric. An important property of the sign-symmetry matrix of a system is that it can be related to the Cauchy index of its transfer function. The Cauchy index of a rational function is defined as follows.

DEFINITION 10.1. Let $p(x)$ and $q(x)$ be relatively prime polynomials with real coefficients. The Cauchy index $C_{\text{ind}}(g(x))$ of $g(x)=p(x)/q(x)$ is defined as the number of jumps from $-\infty$ to $+\infty$ minus the number of jumps from $+\infty$ to $-\infty$ of $g(x)$ as x varies from $-\infty$ to $+\infty$.

A consequence of a result in [2] is that if a system is sign symmetric with respect to a sign-symmetry matrix S, the Cauchy index of its transfer function $g(s)$ is given by

$$C_{\text{ind}}(g(s)) = \text{trace } (S).$$

In [25] it was shown that systems in $C_n^{1,1}$ with Cauchy index n characterize the so-called relaxation systems. These are systems whose impulse response is a completely monotonic function.

Geometric aspects of the parameter space. Another important aspect of the canonical forms presented in this paper is the comparatively simple structure of the parameter space. In many other canonical forms for minimal systems the parameter set at which the systems lose minimality is described by complicated sets of algebraic equations. Here minimality is preserved provided certain parameters are strictly positive. A disadvantage, however, is that even in the case of SISO systems several distinct structures are necessary to parametrize any of the classes of systems considered here.

Let S_n denote any of the following classes of single-input single-output systems: $C_n^{1,1}$, $L_n^{1,1}$, $B_n^{1,1}$, or P_n^1. We have shown that each system in S_n has associated with it a unique set of discrete parameters:

$$n_1, \cdots, n_j, \cdots, n_k, \qquad n_j \in \mathcal{N}, \ \textstyle\sum_{j=1}^k n_j = n,$$
$$s_1, \cdots, s_j, \cdots, s_k, \qquad s_j = \pm 1, \quad 1 \leqq j \leqq k.$$

If we denote by $S_n(n_1, n_2, \cdots, n_k; s_1, s_2, \cdots, s_k)$ the set of systems in S_n with the discrete parameters $(n_1, n_2, \cdots, n_k; s_1, s_2, \cdots, s_k)$, then we can clearly write the set S_n as the disjoint union of the sets $S_n(n_1, n_2, \cdots, n_k; s_1, s_2, \cdots, s_k)$, i.e.,

$$S_n = \bigcup_{\substack{1 \leqq k \leqq n \\ n_1+n_2+\cdots+n_k=n \\ s_1=\pm 1, s_2 \pm 1, \cdots, s_k=\pm 1}} S_n(n_1, n_2, \cdots, n_k; s_1, s_2, \cdots, s_k).$$

It follows easily from the structure of the continuous parameters that for each choice of discrete parameters $n_1, n_2, \cdots, n_k; s_1, s_2, \cdots, s_k$ the parameter set of $S_n(n_1, n_2, \cdots, n_k; s_1, s_2, \cdots, s_k)$ is, in fact, diffeomorphic to $\Re^{n+k} \times \Re$. Therefore we have a decomposition of the set S_n into disjoint "cells," each of which are diffeomorphic to a Euclidean space. The term "cell" is used here in a loose sense and not in its strict topological meaning. We will not go any further into a topological investigation of this decomposition. We refer to [9] and [14] where a different decomposition, which originates from a continued fraction expansion of scalar transfer functions, is introduced and investigated. It is, however, interesting to consider the number of cells in our decomposition. The total number of cells of S_n is easily seen by induction to be $2 \times 3^{n-1}$. The number of cells of dimension $n+k$ is equal to the number of different choices of k blocks of singular values times the possible choices of signs, which is 2^k and therefore gives

$$2^k \binom{n-1}{k-1}.$$

Note that we neglect the parameter that corresponds to the d-term, which is not relevant to the present discussion. Another number of interest is the number of cells of fixed dimension corresponding to a certain Cauchy index. These numbers are not so easily determined and we will just give a small table that contains these numbers for small dimensions.

Number of cells of given dimension and Cauchy index.

Order	dim	Cauchy index								
n	$n+k$	-4	-3	-2	-1	0	1	2	3	4
1	2				1		1			
2	3					2				
	4			1		2		1		
3	4				1		1			
	5				4		4			
	6		1		3		3		1	
4	5					2				
	6			2		8		2		
	7			6		12		6		
	8	1		4		6		4		1

It is surprising to see that the different numbers we have obtained coincide with the numbers found in [9] for the cell decomposition of $L_n^{1,1}$, which was derived from continued fractions of the transfer functions. Simple examples show, however, that this decomposition is not identical with the decomposition derived here. We have not considered minimum-phase systems here. Similar results, however, also hold for these systems.

The geometry of the parameter space was used in [26] to study connectivity properties of the various classes of systems. See [24] for the case of asymptotically stable systems.

REFERENCES

[1] U. M. Al-Saggaf and G. F. Franklin, *An error bound for a discrete reduced order model of a linear multivariable system*, IEEE Trans. Automat. Control, 32 (1987), pp. 815-819.

[2] B. D. O. Anderson, *On the computation of the Cauchy index*, Quart. Appl. Math., 1972, pp. 577-582.

[3] B. D. O. Anderson, K. L. Hitz, and N. D. Diem, *Recursive algorithm for spectral factorization*, IEEE Trans. Circuits and Systems, 21 (1974), pp. 742-750.

[4] B. D. O. Anderson, E. I. Jury, and M. Mansour, *Schwarz matrix properties for continuous and discrete time systems*, Internat. J. Control, 23 (1976), pp. 1-16.

[5] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[6] P. E. Caines, *Linear Stochastic Systems*, John Wiley, New York, 1988.

[7] U. B. Desai and D. Pal, *A transformation approach to stochastic model reduction*, IEEE Trans. Automat. Control, 29 (1984), pp. 1097-1100.

[8] P. L. Faurre, M. Clerget, and F. Germain, *Opérateurs rationnels positifs*, Dunod, Paris, 1979.

[9] P. A. Fuhrmann and P. S. Krishnaprasad, *Towards a cell decomposition for rational functions*, IMA J. Math. Control and Inform., 3 (1986), pp. 137-150.

[10] K. Glover, *All optimal Hankel-norm approximations of linear multivariable systems and their L^∞-error bounds*, Internat. J. Control, 39 (1984), pp. 1115-1193.

[11] K. Glover and J. C. Doyle, *State-space formulae for all stabilizing controllers that satisfy an H_∞-norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167-172.

[12] R. P. GUIDORZI, *Invariants and canonical forms for systems: structural and parametric identification*, Automatica, 17 (1981), pp. 117-133.

[13] U. HELMKE, *Zur Topologie des Raumes linearer Kontrollsysteme*, Ph.D. thesis, University of Bremen, Bremen, Germany, 1982.

[14] U. HELMKE, D. HINRICHSEN, AND W. MANTHEY, *A cell decomposition of the space of real Hankels of rank* $\leqq n$ *and some applications*, Tech. Report 183, Institut für dynamische Systeme, Universität Bremen, Bremen, Germany, 1988.

[15] D. HINRICHSEN, *Canonical forms and parametrization problems in linear systems theory*, in Fourth IMA International Conference on Control Theory, P. A. Cook, ed., Academic Press, New York, 1986.

[16] E. A. JONCKHEERE AND L. M. SILVERMAN, *A new set of invariants for linear systems—application to reduced order compensator design*, IEEE Trans. Automat. Control, 28 (1983), pp. 953-964.

[17] P. T. KABAMBA, *Balanced forms: canonicity and parametrization*, IEEE Trans. Automat. Control, 30 (1985), pp. 1106-1109.

[18] S. S. MAHIL, F. W. FAIRMAN, AND B. S. LEE, *Some integral properties for balanced realizations of scalar systems*, IEEE Trans. Automat. Control, 29 (1984), pp. 181-183.

[19] B. C. MOORE, *Principal component analysis in linear systems: controllability, observability and model reduction*, IEEE Trans. Automat. Control, 26 (1981), pp. 17-32.

[20] R. OBER, *Problems of parametrization of linear systems*, Master's thesis, Engineering Department, Cambridge University, Cambridge, UK, August 1985.

[21] ———, *Asymptotically stable allpass transfer functions: canonical form, parametrization and realization*, in Proc. IFAC World Congress, Munich, 1987.

[22] ———, *Balanced realizations: canonical form, parametrization, model reduction*, Internat. J. Control, 46 (1987), pp. 643-670.

[23] ———, *Balanced realizations for finite and infinite dimensional linear systems*, Ph.D. thesis, Engineering Department, Cambridge University, Cambridge, UK, 1987.

[24] ———, *Topology of the set of asymptotically stable systems*, Internat. J. Control, 46 (1987), pp. 263-280.

[25] ———, *The parametrization of linear systems using balanced realizations: relaxation systems*, in Linear Circuits, Systems and Signal Processing: Theory and Application, C. I. Byrnes, C. F. Martin, and R. E. Saeks, eds., North-Holland, Amsterdam, 1988.

[26] ———, *Connectivity properties of classes of linear systems*, Internat. J. Control, 50 (1989), pp. 2049-2073.

[27] R. OBER AND D. MCFARLANE, *Balanced canonical forms for minimal systems: a normalized coprime factor approach*, Linear Algebra Appl., Special Issue on Linear Systems and Control, 122-124 (1989), pp. 23-64.

[28] P. C. OPDENACKER AND E. A. JONCKHEERE, *A contraction mapping preserving balanced reduction scheme and its infinity norm error bounds*, IEEE Trans. Circuits and Systems, 35 (1988), pp. 184-189.

[29] L. PERNEBO AND L. M. SILVERMAN, *Model reduction via balanced state space representations*, IEEE Trans. Automat. Control, 37 (1982), pp. 382-387.

[30] J. RISSANEN, *Basic of invariants and canonical forms for linear dynamic systems*, Automatica, 10 (1974), pp. 175-182.

[31] M. VIDYASAGAR, *Control System Synthesis: A Factorization Approach*, MIT Press, Cambridge, MA, 1985.

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

[33] D. A. WILSON AND A. KUMAR, *Symmetry properties of balanced systems*, IEEE Trans. Automat. Control, 28 (1983), pp. 927-929.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 6, pp. 1288-1299, November 1991

002

STOCHASTIC REGULATOR THEORY FOR A CLASS OF ABSTRACT WAVE EQUATIONS*

A. V. BALAKRISHNAN†

Abstract. A class of steady-state stochastic regulator problems for abstract wave equations in a Hilbert space—of relevance to the problem of feedback control of large space structures using co-located controls/sensors—is studied. Both the control operator, as well as the observation operator, are finite-dimensional. As a result, the usual condition of exponential stabilizability invoked for existence of solutions to the steady-state Riccati equations is not valid. Fortunately, for the problems considered it turns out that strong stabilizability suffices. In particular, a closed form expression is obtained for the minimal (asymptotic) performance criterion as the control effort is allowed to grow without bound.

Key words. stochastic regulator, abstract wave equation, steady-state Riccati equation

1. Introduction. In this paper we consider a class of steady-state stochastic regulator problems arising in active stabilization of large space structures using co-located sensors and controllers [1]-[4]. We deal only with the abstract formulation of such problems as a wave equation in a Hilbert space (reference may be made to [2] to [3] for specific examples). In this formulation both the control operator and the observation operator are finite-dimensional. As a result, the theory is actually more difficult because the exponential stabilizability condition usually invoked for the existence of solutions of steady-state Riccati equations (e.g., in [4]) is not valid. The latter result—the lack of exponential stabilizability—follows from [7]. On the other hand, the noise model—the driving or "input" noise—is finite-dimensional—so that the complications of the infinite-dimensional Wiener process (see [5]) are avoided—even though, of course, it induces infinite-dimensional noise in the state variables. In particular, it turns out that steady-state covariance operators need not be nuclear, so that white noise theory [6] is more convenient (even if not essential, since the systems considered are linear!). Some background in this context is provided in [6]. Also we make extensive use of the notions and results of stochastic control theory developed in [6].

We begin in § 2 with the basic system description and problem formulation. The properties of the system essential in the sequel are outlined. In particular, the result of Benchimol [9] on strong stability plays a crucial role throughout. In fact, all "closed-loop" system semigroups are strongly stable and turn out to be adequate for asymptotic stationarity of the processes generated.

Section 3 presents the main results of proofs. A useful corollary provides an expression for the minimal attainable performance as control effort is increased without bound—in particular, in the application context it yields the minimal attainable mean square pointing error using co-located rate and attitude sensors.

2. System description and problem formulation. We begin with the state equation, which is a wave equation in a Hilbert space with random noise input:

$$M\ddot{x}(t)+Ax(t)+Bu(t)+BN_a(t)=0. \tag{2.1}$$

* Received by the editors February 7, 1990; accepted for publication (in revised form) November 30, 1990. This research was supported in part by National Aeronautics and Space Administration task assignment 49: NAS1-18585.

† Electrical Engineering Department, University of California, 6731 Boelter Hall, # 159410, Los Angeles, California 90024.

With $\mathscr{H}$ denoting the (separable) Hilbert space, M is a linear bounded self-adjoint operator on $\mathscr{H}$ into $\mathscr{H}$ which is nonnegative definite with a bounded inverse (zero is in the resolvent of M). In what follows we may, and will, assume M is the identity, without loss of generality in the theory. The operator A is unbounded, closed, with domain dense in $\mathscr{H}$ and is self-adjoint nonnegative definite, and has a compact resolvent with zero in the resolvent set. The operator B is finite-dimensional mapping $\mathscr{H}$ into R^n. The random noise is white Gaussian with spectral density matrix D_a. The control input $u(\cdot)$ is to be physically realizable so that $u(t)$ at time t is to be based only on observed data up to time t. The observations $v(\cdot)$ is given by

$$v(t) = \begin{vmatrix} v_p(t) \\ v_r(t) \end{vmatrix},$$

where

$$v_p(t) = B^*x(t) + N_p(t),$$
$$v_r(t) = B^*\dot{x}(t) + N_r(t), \tag{2.2}$$

where $N_p(\cdot)$, $N_r(\cdot)$ are independent white noises (independent also of the state noise $N_a(\cdot)$) with nonsingular spectral density matrices D_p and D_r, respectively. The optimization problem is a "stochastic regulator" problem: Minimize

$$\lim_{T\to\infty} \frac{1}{T}\left[\int_0^T \|B^*\dot{x}(t)\|^2\,dt + \lambda \int_0^T \|u(t)\|^2\,dt\right] \tag{2.3}$$

for fixed $\lambda > 0$. We are also interested in what happens as $\lambda \to 0$.

We begin by recasting (2.1) in the usual way with a slight difference in that a special inner product is introduced, which in fact is crucial in what follows.

Thus let $\mathscr{H}_E$ denote the energy-inner product space

$$\mathscr{H}_E = \mathscr{D}(\sqrt{A}) \times \mathscr{H}$$

with inner product defined by

$$[Y, Z]_E = [\sqrt{A}\,y_1, \sqrt{A}\,z_1] + [y_2, z_2],$$

where

$$Y = \begin{vmatrix} y_1 \\ y_2 \end{vmatrix}, \qquad Z = \begin{vmatrix} z_1 \\ z_2 \end{vmatrix}.$$

Defining, in the usual fashion (see [2], [6]),

$$\mathscr{A} = \begin{bmatrix} 0 & I \\ -A & 0 \end{bmatrix}, \qquad \mathscr{D}(\mathscr{A}) = \mathscr{D}(A) \times \mathscr{H},$$

$$\mathscr{B}u = \begin{bmatrix} 0 \\ -Bu \end{bmatrix}, \qquad \mathscr{B} \text{ mapping } R^n \text{ into } \mathscr{H}_E,$$

we being with (2.1) in the form

$$\dot{Y}(t) = \mathscr{A}Y(t) + \mathscr{B}u(t) + \mathscr{B}N_a(t), \tag{2.4}$$

where we note that (the adjoint $\mathscr{A}^*$ in $\mathscr{H}_E$ (in the energy-inner product!)):

$$\mathscr{A} + \mathscr{A}^* = 0, \qquad \mathscr{D}(\mathscr{A}) = \mathscr{D}(\mathscr{A}^*)$$

and $\mathscr{A}$ generates a strongly continuous isometric semigroup (actually a group). Also the observation $v(\cdot)$ goes over as

$$v(t) = CY(t) + N_0(t), \tag{2.5}$$

where

$$CY = \begin{vmatrix} B^*y_1 \\ B^*y_2 \end{vmatrix}, \qquad N_0(t) = \begin{vmatrix} N_p(t) \\ N_r(t) \end{vmatrix},$$

where C maps $\mathscr{H}_E$ into $R^n \times R^n$.

Next we make the crucial assumption that $\mathscr{A} \sim \mathscr{B}$ is controllable. A sufficient condition in terms of A and B is this: let ϕ be any eigenvector A:

$$A\phi = \omega^2\phi.$$

Then

$$B^*\phi = 0$$

only if $\phi = 0$. The proof exploits the fact that the eigenvectors of A are complete in $\mathscr{H}$. The controllability condition yields the Benchimol [9] result, which plays a central role in our theory that

$$\mathscr{A} - \mathscr{B}\mathscr{B}^*$$

is strongly stable. Furthermore, with $S_1(\cdot)$ denoting the semigroup generated by $(\mathscr{A} - \mathscr{B}\mathscr{B}^*)$ we have that for $Y \in \mathscr{D}(\mathscr{A})$,

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\|S_1(t)Y\|^2 &= \frac{1}{2}\{((\mathscr{A} - \mathscr{B}\mathscr{B}^*)S_1(t)Y, S_1(t)Y)_E + (S_1(t)Y, (\mathscr{A} - \mathscr{B}\mathscr{B}^*)S_1(t)Y)_E\} \\ &= -\|\mathscr{B}^*S_1(t)Y\|^2,\end{aligned}$$

and hence

$$\frac{1}{2}\|Y\|_E^2 - \frac{1}{2}\|S_1(t)Y\|_E^2 = \int_0^t \|\mathscr{B}^*S_1(\sigma)Y\|^2\,d\sigma.$$

The domain of $\mathscr{A}$ being dense in $\mathscr{H}_E$, this holds for every Y in $\mathscr{H}_E$. Hence taking limits as $t \to \infty$ we obtain

$$\frac{\|Y\|_e^2}{2} = \int_0^\infty \|\mathscr{B}^*S_1(\sigma)\|^2\,d\sigma. \tag{2.6}$$

Replacing $\mathscr{A}$ by $\mathscr{A}^*$ and equivalently $S_1(\cdot)$ by $S_1(\cdot)^*$, we also have that

$$\frac{\|Y\|_E^2}{2} = \int_0^\infty \|\mathscr{B}^*S_1(\sigma)^*Y\|^2\,d\sigma. \tag{2.6a}$$

Finally, (2.3) becomes: Minimize

$$\lim_{T\to\infty}\left\{\frac{1}{T}\int_0^T \|\mathscr{B}^*Y(t)\|^2\,dt + \frac{\lambda}{T}\int_0^T \|u(t)\|^2\,dt\right\} \tag{2.7}$$

(where $\mathcal{B}^*$ is the adjoint of $\mathcal{B}$ in the energy-inner product!) We note here, for future reference, that

$$\mathcal{L}^* v_p = \left|\begin{matrix} A^{-1}Bv_p \\ 0 \end{matrix}\right| \quad \text{if } \mathcal{L}Y = B^* y_1, \tag{2.8}$$

$$\mathcal{B}^* Y = -B^* y_2, \tag{2.9}$$

$$C^* v = \left|\begin{matrix} A^{-1}Bv_p \\ Bv_r \end{matrix}\right| \tag{2.10}$$

(where C^* is the adjoint of C in $\mathcal{H}_E$ (in the energy-inner products),

$$C^* D_0^{-1} C Y = \left|\begin{matrix} A^{-1}BD_p^{-1}B^* y_1 \\ BD_r^{-1}B^* y_2 \end{matrix}\right|, \tag{2.11}$$

where D_0 is the spectral density matrix of the process $N_0(\cdot)$).

The stochastic regulator problem is then embodied in (2.4), (2.5), and (2.7).

3. Main results.

THEOREM 3.1. *The stochastic regulator problem* (2.4), (2.5), (2.7) *has the optimal solution given by*

$$u_0(t) = \frac{-\mathcal{B}^* P_c \hat{Y}(t)}{\lambda}, \tag{3.1}$$

where P_c is the unique self-adjoint solution of

$$0 = [P_c Y, \mathcal{A} Y] + [\mathcal{A} Y, P_c Y] + [\mathcal{B}^* Y, \mathcal{B}^* Y] - \frac{[\mathcal{B}^* P_c Y, \mathcal{B}^* P_c Y]}{\lambda}, \qquad Y \in \mathcal{D}(\mathcal{A}), \tag{3.2}$$

and $\hat{Y}(\cdot)$ is defined by the "Kalman filter" equation

$$\dot{\hat{Y}}(t) = (\mathcal{A} - P_f C^* D_0^{-1} C)\hat{Y}(t) - \frac{\mathcal{B}\mathcal{B}^*}{\lambda} P_c \hat{Y}(t) + P_f C^* D_0^{-1} v(t), \tag{3.3}$$

where P_f is the unique self-adjoint solution of

$$\begin{aligned} 0 = {} & [P_f Y, \mathcal{A}^* Y] + [\mathcal{A}^* Y, P_f Y] + [\mathcal{B} D_a \mathcal{B}^* Y, Y] \\ & -[P_f C^* D_0^{-1} C P_f Y, Y], \qquad Y \in \mathcal{D}(\mathcal{A}^*). \end{aligned} \tag{3.4}$$

The corresponding (minimal) cost functional is

$$= \operatorname{Tr} \mathcal{B}^* P_f \mathcal{B}^* + \operatorname{Tr} \sqrt{D_0^{-1}}\, C P_f P_c P_f C^* \sqrt{D_0^{-1}}. \tag{3.5}$$

Also

$$\lim_{T \to \infty} \frac{1}{T} \int_0^T [\mathcal{B}^* Y(t), \mathcal{B}^* Y(t)]\, dt$$

goes (decreases) to

$$\operatorname{Tr} \mathcal{B}^* P_f \mathcal{B} \tag{3.6}$$

as $\lambda \to 0$.

We begin with some preliminary lemmas.

LEMMA 1. $\mathcal{A} - C^* D_0^{-1} C$ *generates a strongly stable semigroup $S_0(\cdot)$ such that for every Y in $\mathcal{H}_E$*

$$\int_0^\infty \|\sqrt{D_0^{-1}}\, C S_0(\sigma) Y\|^2 \, d\sigma < \infty. \tag{3.7}$$

Proof. Since $\mathscr{A}$ generates an isometric group it is enough to show that

$$\mathscr{A} \sim C^*\sqrt{D_0^{-1}}$$

is controllable. Let $S(\cdot)$ denote the semigroup generated by $\mathscr{A}$. Suppose

$$\sqrt{D_0^{-1}}\, CS(t)^* Y = 0, \qquad t \geqq 0. \tag{3.8}$$

We need to show that Y must then be zero. Now letting

$$S(t)^* Y = \begin{vmatrix} y_1(t) \\ y_2(t) \end{vmatrix}$$

we have that (since D_0 is nonsingular)

$$CS(t)^* Y = \begin{vmatrix} B^* y_1(t) \\ B^* y_2(t) \end{vmatrix} = 0, \qquad t \geqq 0. \tag{3.9}$$

However,

$$B^* y_2(t) = -\mathscr{B}^* S(t)^* Y$$

and, in particular, (3.9) implies that

$$\mathscr{B}^* S(t)^* Y = 0, \qquad 0 \leqq t,$$

which because of our assumption that

$$(\mathscr{A} \sim \mathscr{B})$$

is controllable, implies that

$$Y = 0.$$

Hence $(\mathscr{A} \sim C^*\sqrt{D_0^{-1}})$ is controllable and $(\mathscr{A} - C^* D_0^{-1} C)$ is strongly stable. Hence also (3.7) follows from (2.6) taking $\mathscr{B} = C^*\sqrt{D_0^{-1}}$ therein. Also using $\mathscr{A}^*$ in place of $\mathscr{A}$, we may replace $S_0(\sigma)$ by $S_0(\sigma)^*$.

As a consequence of this lemma we can state Lemma 2.

LEMMA 2. *Let the control $u(t) \equiv 0$, and define the process $\hat{Y}(\cdot)$ by*

$$\dot{\hat{Y}}(t) = (\mathscr{A} - \gamma C^* D_0^{-1} C)\hat{Y} + \gamma C^* D_0^{-1} v, \qquad \gamma > 0.$$

Then we have

$$\dot{Y}(t) = \mathscr{A} Y(t) + \mathscr{B} N_a(t),$$

$$\dot{\hat{Y}}(t) = \mathscr{A}\hat{Y}(t) + \gamma C^* D_0^{-1} C(Y(t) - \hat{Y}(t)) + \gamma C^* D_0^{-1} N_0,$$

and

$$Z(t) = Y(t) - \hat{Y}(t)$$

satisfies

$$\dot{Z}(t) = (\mathscr{A} - \gamma C^* D_0^{-1} C) Z(t) + \gamma C^* D_0^{-1} N_0 + \mathscr{B} N_a(t).$$

Let

$$\mathscr{R}(t) = E[Z(t) Z(t)^*].$$

Then $R(t)$ converges strongly to $R(\infty)$ and

$$[R(\infty) Y, Y] = \frac{\gamma \|Y\|_E^2}{2} + \int_0^\infty \|\sqrt{D_a}\, \mathscr{B}^* S_\gamma(s)^* Y\|^2 \, d\sigma, \tag{3.10}$$

where $S_\gamma(\cdot)$ is the semigroup generated by $\mathcal{A}-\gamma C^*D_0^{-1}C$, and the second term

$$\int_0^\infty \|\sqrt{D_a}\,\mathcal{B}^*S_\gamma(\sigma)Y\|^2\,d\sigma \leqq \frac{\|Y\|_E^2}{2\gamma}\|D_a\|\,\|D_0\|.$$

Proof. The proof is fairly immediate from

$$[R(t)Y, Y]=[\Lambda S_\gamma(t)^*Y, S_\gamma(t)^*Y] + \int_0^t [(\gamma^2C^*D_0^{-1}C+\mathcal{B}D_a\mathcal{B}^*)S_\gamma^*(\sigma)Y, S_\gamma^*(\sigma)Y]\,d\sigma,$$

where

$$\Lambda = E[(Y(0)-\hat{Y}(0))(Y(0)-\hat{Y}(0))^*].$$

See also [6] if necessary for a similar argument. Also

$$\int_0^\infty \|\mathcal{B}^*S_\gamma^*(\sigma)Y\|^2\,d\sigma<\infty$$

follows from

$$\int_0^\infty \|\mathcal{B}^*S_\gamma^*(\sigma)Y\|^2\,d\sigma \leqq \int_0^\infty \|CS_\gamma^*(\sigma)Y\|^2\,d\sigma \leqq \int_0^\infty \|D_0\|\,\|\sqrt{D_0^{-1}}\,CS_\gamma^*(\sigma)Y\|^2\,d\sigma = \frac{\|D_0\|\,\|Y\|_E^2}{2\gamma}.$$

THEOREM 3.2. *There exist admissible ("feedback") controls for which* (2.7) *is finite.*

Proof. We ignore $v_p(\cdot)$ and devise controls using only $v_r(\cdot)$ and the results in [7]. Note that

$$v_r(t)=-\mathcal{B}^*Y(t)+N_r(t).$$

If $D_a=d_aI$ and $D_0=d_0I$, we could use the results in [7] verbatim. For the more general case considered here we will now show that the choice

$$u(t)=-\mathcal{B}^*\hat{Y}(t), \tag{3.11}$$

where

$$\begin{aligned}\dot{\hat{Y}}(t)&=\mathcal{A}\hat{Y}(t)-\mathcal{B}(v_r(t)+\mathcal{B}^*\hat{Y}(t))-\mathcal{B}\mathcal{B}^*\hat{Y}(t),\\ \hat{Y}(0)&=0\end{aligned} \tag{3.12}$$

yields an admissible control which makes (2.7) finite. We may think of $\hat{Y}(\cdot)$ defined by (3.12) as a "suboptimal" estimate, and, of course, (3.11) is a suboptimal control—in other words, this is merely a technique to show that admissible controls exist for which (2.7) is finite.

LEMMA 3. *The operator*

$$\mathcal{A}_2=\begin{vmatrix}\mathcal{A}^*-\mathcal{B}\mathcal{B}^* & \mathcal{B}\mathcal{B}^*\\ \mathcal{B}\mathcal{B}^* & \mathcal{A}^*-2\mathcal{B}\mathcal{B}^*\end{vmatrix} \tag{3.13}$$

with domain $(\mathcal{D}(\mathcal{A})\times\mathcal{D}(\mathcal{A}))$ in $H_E\times H_E$ generates a strongly continuous semigroup

$S_Z(\cdot)$, which is strongly stable. Moreover, using the notation

$$S_Z(t)Z = \begin{vmatrix} Y(t) \\ \hat{Y}(t) \end{vmatrix}, \qquad Z = \begin{vmatrix} Y \\ \hat{Y} \end{vmatrix}$$

we have further that

$$\int_0^\infty \|\mathscr{B}^* Y(t)\|^2 \, dt + \int_0^\infty \|\mathscr{B}^* \hat{Y}(t)\|^2 \, dt < \infty \tag{3.14}$$

for each Z.

Proof. The semigroup generation property is immediate; only the strong stability property and (3.14) require additional proof.

Let $Z \in \mathscr{D}(\mathscr{A}_Z)$. Then if

$$S_Z(t)Z = \begin{vmatrix} Y(t) \\ \hat{Y}(t) \end{vmatrix},$$

we have

$$\dot{Y}(t) + \dot{\hat{Y}}(t) = (\mathscr{A}^* - \mathscr{B}\mathscr{B}^*)(Y(t) + \hat{Y}(t)).$$

Hence

$$Y(t) + \hat{Y}(t) = S_1(t)(Y + \hat{Y}), \qquad t \geqq 0, \tag{3.15}$$

where $S_1(\cdot)$ is the semigroup generated by $(\mathscr{A}^* - \mathscr{B}\mathscr{B}^*)$, and (3.15) continues to hold for all Z in $\mathscr{H}_E \times \mathscr{H}_E$. Similarly, from

$$\dot{\hat{Y}}(t) = \mathscr{A}^* Y(t) + \mathscr{B}\mathscr{B}^* \hat{Y}(t) = (\mathscr{A}^* - \mathscr{B}\mathscr{B}^*) Y(t) + \mathscr{B}\mathscr{B}^*(Y(t) + \hat{Y}(t)),$$

we have

$$Y(t) = S_1(t) Y + \int_0^t S_1(t-\sigma)\mathscr{B}\mathscr{B}^* S_1(\sigma)(Y + \hat{Y}) \, d\sigma \tag{3.16}$$

and also, similarly,

$$\hat{Y}(t) = S_1(t) \hat{Y} + \int_0^t S_1(t-\sigma)\mathscr{B}\mathscr{B}^* S_1(\sigma)(Y + \hat{Y}) \, d\sigma \tag{3.17}$$

for all Y, $\hat{Y}$ in $\mathscr{H}_E$. For any ψ in $\mathscr{H}_E$, let us show now that

$$[Y(t), \psi] \to 0 \quad \text{as } t \to \infty.$$

We have

$$[Y(t), \psi] = [S_1(t) Y, \psi] + \int_0^t [\mathscr{B}^* S_1(\sigma)(Y + \hat{Y}), \mathscr{B}^* S_1(t-\sigma)\psi] \, d\sigma. \tag{3.18}$$

Now $S_1(\cdot)$ being strongly stable and

$$\int_0^\infty \|\mathscr{B}^* S_1(\sigma) X\|^2 \, d\sigma < \infty \quad \text{for any } X \text{ in } \mathscr{H}_E$$

it follows readily that (3.18) goes to zero as $t \to \infty$. Similarly, so does $[\hat{Y}(t), \psi]$ using (3.17). Hence the semigroup $S_Z(\cdot)$ is weakly stable, and since it has a compact resolvent it follows that it is actually strongly stable (see [6] if necessary). Next

$$\mathscr{B}^* Y(t) = \mathscr{B}^* S_1(t) Y + \int_0^t \mathscr{B}^* S_1(t-\sigma)\mathscr{B}\mathscr{B}^* S_1(\sigma)(Y + \hat{Y}) \, d\sigma \tag{3.19}$$

since

$$\int_0^\infty \|\mathcal{B}^* S_1(t) Y\|^2 \, dt < \infty$$

and the second term is a convolution of functions in $L_2[0, \infty]$,

$$\int_0^\infty \|\mathcal{B}^* Y(t)\|^2 \, dt < \infty.$$

By a similar argument, using (3.17),

$$\int_0^\infty \|\mathcal{B}^* \hat{Y}(t)\|^2 \, dt < \infty$$

and hence we have proved (3.14).

We can now proceed to prove Theorem 3.2. We consider the coupled equations

$$\begin{aligned} \dot{Y}(t) &= \mathcal{A}Y(t) - \mathcal{B}\mathcal{B}^* \hat{Y}(t) + \mathcal{B}N_a(t), \\ \dot{\hat{Y}}(t) &= (\mathcal{A} - 2\mathcal{B}\mathcal{B}^*)\hat{Y}(t) + \mathcal{B}\mathcal{B}^* Y(t) - \mathcal{B}N_r(t). \end{aligned} \tag{3.20}$$

Writing

$$Z(t) = \left| \begin{matrix} Y(t) \\ \hat{Y}(t) \end{matrix} \right|, \qquad Z = \left| \begin{matrix} Y \\ \hat{Y} \end{matrix} \right|,$$

we obtain

$$\dot{Z}(t) = A_Z Z(t) + \left| \begin{matrix} \mathcal{B}N_a(t) \\ \mathcal{B}(-N_r(t)) \end{matrix} \right|.$$

Let

$$E[Z(t)Z(t)^*] = \mathcal{R}(t), \qquad \mathcal{R}(0) = \Lambda,$$

where, of course, $\mathcal{R}(t)$, $t \geqq 0$ is a linear bounded operator mapping $\mathcal{H}_E \times \mathcal{H}_E$ into itself, and Λ is self-adjoint, nonnegative definite and so is $\mathcal{R}(t)$ for each t. Then

$$[\mathcal{R}(t)Z, Z] = [\Lambda S_Z(t)Z, S_Z(t)Z] + \int_0^t \left[\left| \begin{matrix} \mathcal{B}D_a\mathcal{B}^* & 0 \\ 0 & \mathcal{B}D_r\mathcal{B}^* \end{matrix} \right| S_Z(\sigma)Z, S_Z(\sigma)Z \right] d\sigma,$$

where $S_Z(\cdot)$ is the adjoint of the semigroup generated by

$$\left| \begin{matrix} \mathcal{A} - \mathcal{B}\mathcal{B}^* & -\mathcal{B}\mathcal{B}^* \\ \mathcal{B}\mathcal{B}^* & \mathcal{A} - 2\mathcal{B}\mathcal{B}^* \end{matrix} \right|.$$

By Lemma 3, $S_Z(\cdot)$ is strongly stable and the second term in

$$\int_0^\infty [\mathcal{B}D_a\mathcal{B}^* Y(\sigma), Y(\sigma)] \, d\sigma + \int_0^\infty [\mathcal{B}D_r\mathcal{B}^* \hat{Y}(\sigma), \hat{Y}(\sigma)] \, d\sigma$$

(where we use the notation $S_Z(\sigma)Z = \left|{Y(\sigma) \atop \hat{Y}(\sigma)}\right|$)

$$\leqq \int_0^\infty \|D_a\| \, \|\mathcal{B}^* Y(\sigma)\|^2 \, d\sigma + \|D_a\| \int_0^\infty \|\mathcal{B}^* \hat{Y}(\sigma)\|^2 \, d\sigma$$

$$< \infty$$

by (3.14). Hence

$$\mathscr{R}(t) \text{ converges strongly to } \mathscr{R}(\infty),$$

$$[\mathscr{R}(\infty)Z, Z] < \infty \quad \text{for every } Z \text{ in } \mathscr{H}_E \times H_E,$$

and $\mathscr{R}(\infty)$ is linear bounded, self-adjoint, and nonnegative definite. Hence the process $Z(\cdot)$ defined (3.20) has a steady state, with steady-state covariance

$$\lim_{T\to\infty} E[Z(T)Z(T+t)^*] = S_Z(t)\mathscr{R}(\infty).$$

In particular, the finite-dimensional process

$$\mathscr{B}^* Y(t)$$

is Gaussian and asymptotically stationary. Therefore

$$(3.21) \qquad \lim_{T\to\infty} \frac{1}{T}\int_0^T \|\mathscr{B}^* Y(t)\|^2\, dt = \operatorname{Tr} \mathscr{B}^*\mathscr{R}_{11}\mathscr{B} < \infty$$

and, similarly,

$$(3.22) \quad \lim_{T\to\infty} \frac{1}{T}\int_0^T \|u(t)\|^2\, dt = \lim_{T\to\infty} \frac{1}{T}\int_0^T \|\mathscr{B}^* \hat{Y}(t)\|^2\, dt = \operatorname{Tr} \mathscr{B}^*\mathscr{R}_{22}B < \infty,$$

where we have used the representation for $\mathscr{R}(\infty)$:

$$\mathscr{R}(\infty) = \begin{vmatrix} \mathscr{R}_{11} & \mathscr{R}_{12} \\ \mathscr{R}_{21} & \mathscr{R}_{22} \end{vmatrix},$$

$\mathscr{R}_{ij}$ being linear bounded operators mapping $\mathscr{H}_E$ into $\mathscr{H}_E$, thus proving Theorem 3.2.

We are now ready to prove our main results in Theorem 3.1.

We follow in the main the arguments in [6, p. 351 et seq.], except that now we do not invoke exponential stability—replacing it, in fact, by the strong stability property. Thus let $P_f(t)$, $t \geqq 0$ be the solution of (the Riccati equation):

$$(3.23) \quad \frac{d}{dt}[P_f(t)Y, Y] = [\mathscr{A}^*Y, P_f(t)Y] + [P_f(t)Y, \mathscr{A}^*Y] + [\mathscr{B}D_a\mathscr{B}^*Y, Y]$$
$$-[P_fC^*D_0^{-1}CP_fY, Y]$$

for every Y in $\mathscr{D}(\mathscr{A}^*)(=\mathscr{D}(\mathscr{A}))$, with

$$P_f(0) = 0.$$

See [6] for the existence and uniqueness of solution. $P_f(t)$, $t \geqq 0$, is a self-adjoint, nonnegative definite mapping $\mathscr{H}_E$ into $\mathscr{H}_E$. Similarly, let $P_c(t)$ be the solution of

$$(3.24) \quad \frac{d}{dt}[P_c(t)Y, Y] = [P_c(t)Y, \mathscr{A}Y] + [\mathscr{A}Y, P_c(t)Y] + [\mathscr{B}^*Y, \mathscr{B}^*Y]$$
$$-\frac{[\mathscr{B}^*P_c(t)Y, \mathscr{B}^*P_c(t)Y]}{\lambda}$$

for Y in $\mathscr{D}(\mathscr{A})$, and

$$P_c(0) = 0.$$

Then both

$$[P_f(t)Y, Y] \quad \text{and} \quad [P_c(t)Y, Y]$$

are monotone nondecreasing, and our first step is to show that the limit is finite in each case. For this purpose let us use the control defined in Theorem 3.2. Then

$$\begin{aligned}(3.25)\qquad &\frac{1}{T}\int_0^T E[\mathscr{B}^*Y(t),\mathscr{B}^*Y(t)]\,dt+\frac{1}{T}\int_0^T E(\|u(t)\|^2)\,dt\\ &\geqq\frac{1}{T}\int_0^T \operatorname{Tr}\mathscr{B}^*P_f(t)\mathscr{B}^*\,dt\\ &+\frac{1}{T}\int_0^T \operatorname{Tr}\sqrt{D_0^{-1}}\,CP_f(t)P_c(T-t)P_f(t)C^*\sqrt{D_0^{-1}}\,dt.\end{aligned}$$

Now by Lemma 2, we see that

$$[P_f(t)Y,Y]<[R(\infty)Y,Y]<\infty$$

in the notation of (3.10). Hence $P_f(t)$ converges strongly to P_f, say. Then the right side of (3.23) converges to the right side of (3.4), while the left side of (3.23) since it converges must converge to zero. Hence P_f satisfies (3.4), and is, in fact, the unique self-adjoint linear bounded operator solution of (3.4). Next let us consider (3.24). Now we know that

$$[P_c(T)Y,Y]\leqq\int_0^T\|\mathscr{B}^*Y(t)\|^2\,dt+\lambda\int_0^T\|u(t)\|^2\,dt,$$

where

$$\dot Y(t)=\mathscr{A}Y+\mathscr{B}u,\qquad Y(0)=Y$$

for any $u(\cdot)$ in $L_2[0,T]$. Take now

$$u(t)=-\mathscr{B}^*Y(t).$$

Then

$$\begin{aligned}[P_c(T)Y,Y]&\leqq\int_0^T\|\mathscr{B}^*S(t)Y\|^2\,dt+\lambda\int_0^T\|\mathscr{B}^*S(t)Y\|^2\,dt\\ &\leqq\int_0^\infty\|\mathscr{B}^*S(t)Y\|^2\,dt+\lambda\int_0^\infty\|\mathscr{B}^*S(t)Y\|^2\,dt<\infty.\end{aligned}$$

Hence $[P_c(t)Y,Y]$ increases to a finite limit, or $P_c(\infty)$ (self-adjoint linear bounded) which then, of course, satisfies (3.2). In fact, it is immediate that

$$(3.26)\qquad P_c=P_c(\infty)=\sqrt{\lambda}\,I.$$

Hence the right side of (3.25) converges to

$$\operatorname{Tr}\mathscr{B}^*P_f\mathscr{B}+\operatorname{Tr}\sqrt{D_0^{-1}}\,CP_fP_cP_fC^*\sqrt{D_0^{-1}},$$

which is (3.5). Moreover, the choice (3.1) yields the cost functional (3.5). In fact, the system

$$(3.27)\qquad \dot{\hat Y}(t)=(\mathscr{A}-P_fC^*D_0^{-1}C)\hat Y(t)-\frac{\mathscr{B}\mathscr{B}^*}{\sqrt{\lambda}}\hat Y(t)+P_fC^*D_0^{-1}v(t),$$

$$(3.28)\qquad \dot Y(t)=\mathscr{A}Y(t)-\frac{\mathscr{B}\mathscr{B}^*}{\sqrt{\lambda}}\hat Y(t)+BN_a(t)$$

is asymptotically stationary, and

$$\lim_{t\to\infty} E[(Y(t) - \hat{Y}(t))Y(t)^*] = 0. \tag{3.29}$$

It is more convenient to replace (3.27), (3.28) by the system

$$\begin{matrix}\hat{Y}(t) = \\ Z(t) = \end{matrix}\begin{vmatrix} \mathscr{A} - \mathscr{B}\mathscr{B}^*/\lambda & P_f C^* D_0^{-1} C \\ 0 & \mathscr{A} - P_f C^* D_0^{-1} C \end{vmatrix}\begin{vmatrix} \hat{Y}(t) \\ Z(t) \end{vmatrix} + \begin{vmatrix} P_f C^* \sqrt{D_0^{-1}}\, N_0 \\ \mathscr{B} N_a(t) + P_f C^* \sqrt{D_0^{-1}}\, N_0(t) \end{vmatrix}, \tag{3.30}$$

where

$$Z(t) = Y(t) - \hat{Y}(t). \tag{3.31}$$

We can verify directly that the semigroup generated by

$$\begin{vmatrix} \mathscr{A} - \mathscr{B}\mathscr{B}^*/\lambda & P_f C^* D_0^{-1} C \\ 0 & \mathscr{A} - P_f C^* D_0^{-1} C \end{vmatrix}$$

is strongly stable and, furthermore, that system (3.30) is asymptotically stationary. Let

$$\mathscr{R} = \begin{vmatrix} \mathscr{R}_{11} & \mathscr{R}_{12} \\ \mathscr{R}_{21} & \mathscr{R}_{22} \end{vmatrix}$$

denote the steady-state covariance. Then we have, of course,

$$\mathscr{R}_{22} = P_f, \qquad \mathscr{R}_{12} = \mathscr{R}_{22} = 0$$

and $\mathscr{R}_{11}$is defined by

$$\left[\mathscr{R}_{11} Y, \left(\mathscr{A}^* - \frac{\mathscr{B}\mathscr{B}^*}{\lambda}\right) Y\right] + \left[\left(\mathscr{A}^* - \frac{\mathscr{B}\mathscr{B}^*}{\lambda}\right) Y, \mathscr{R}_{11} Y\right] + [P_f C^* D_0^{-1} C P_f Y, Y] = 0, \tag{3.32}$$

$$Y \in \mathscr{D}(\mathscr{A}^*).$$

An important property of $\mathscr{R}_{11}$ that follows from (3.32) is that $\mathscr{R}_{11}$ goes to zero strongly as $\lambda \to 0$. In particular, we have that

$$\lim_{\lambda\to 0}\left[\lim_{T\to\infty} \frac{1}{T}\int_0^T \|\mathscr{L}Y(t)\|^2\, dt\right] = \operatorname{Tr} \mathscr{L}P_f\mathscr{L}^*. \tag{3.33}$$

As a result we have the following useful corollary.

COROLLARY. *Consider the stochastic regulator problem for the system* (2.4), (2.5) *with the cost functional*

$$\lim_{T\to\infty} \left\langle \frac{1}{T}\int_0^T \|\mathscr{L}Y(t)\|^2\, dt + \frac{\lambda}{T}\int_0^T \|u(t)\|^2\, dt\right] \tag{3.34}$$

(*replacing* $\mathscr{B}^*$ *in* (2.7) *by* $\mathscr{L}$).

Let $q(\lambda)$ *denote the infimum of* (3.34). *Then*

$$\lim_{T\to 0} q(\lambda) = \operatorname{Tr} \mathscr{L}P_f\mathscr{L}^*. \tag{3.35}$$

In particular,

$$\lim_{\lambda\to 0}\left[\frac{1}{T}\int_0^T \|\mathscr{L}Y(t)\|^2\, dt\right] \geqq \operatorname{Tr} \mathscr{L}P_f\mathscr{L}^*. \tag{3.36}$$

Remark. The novelty in this result is that we do not need to exhibit the optimal control that minimizes (3.34) for each λ; and indeed we need not thus discuss whether an optimal control exists! Of course, $\mathscr{L}$ can be replaced by any finite-dimensional operator on $\mathscr{H}_E$.

Proof. For each $\lambda > 0$, the control $u(\cdot)$ defined by (3.1) which is optimal for (2.7), is certainly an admissible control since the corresponding value for (3.34) is

$$= \mathrm{Tr}\, \mathscr{L} P_f \mathscr{L}^* + \mathrm{Tr}\, \mathscr{L} \mathscr{R}_{11} \mathscr{L}^* + \mathrm{Tr}\, \mathscr{B}^* \mathscr{R}_{11} \mathscr{B}. \tag{3.37}$$

Hence

$$\lim_{\lambda \to 0} q(\lambda) \leqq \mathrm{Tr}\, \mathscr{L} P_f \mathscr{L}^*.$$

On the other hand, for any admissible control $u(\cdot)$

$$\lim_{t \to \infty} E[\mathscr{L} Y(t) Y(t)^* \mathscr{L}^*] = \mathrm{Tr}\, \mathscr{L} P_f \mathscr{L}^* + \mathrm{Tr}\, \mathscr{L} \hat{R} \mathscr{L}^*,$$

where

$$\hat{R} = \lim_{t \to \infty} E[\hat{Y}(t) \hat{Y}(t)^*],$$

where

$$\hat{Y}(t) = E[Y(t) \,|\, v(s), s \leqq t].$$

Hence

$$q(\lambda) \geqq \mathrm{Tr}\, \mathscr{L} P_f \mathscr{L}^*$$

and hence the result follows. Note, in particular, that

$$\mathscr{L} P_f \mathscr{L}^* = \mathrm{Tr}\, B^* P_{11} A^{-1} B,$$

where

$$P_f = \begin{vmatrix} P_{11} & P_{22} \\ P_{21} & P_{22} \end{vmatrix}.$$

REFERENCES

[1] A. V. Balakrishnan, *A mathematical formulation of the* SCOLE *control problem*, NASA CR 172581, May 1985.

[2] ———, *On a Large Space Structure Control Problem*, Lecture Notes in Control and Information Sciences, Vol. 97, Springer-Verlag, New York, 1987, pp. 3–15.

[3] ———, *Control of Flexible Flight Structures*, in Analyse Mathematique et Applications, Gauthier-Villars, Paris, 1988.

[4] J. Zabczyk, *Remarks on the algebraic Riccati equation in Hilbert space*, J. Appl. Math. Optim., 2 (1975/76), pp. 251–258.

[5] K. Ito, *Foundations of Stochastic Differential Equations in Infinite Dimensional Spaces*, Society for Industrial and Applied Mathematics, Philadelphia, PA, 1984.

[6] A. V. Balakrishnan, *Applied Functional Analysis*, 2nd ed., Springer-Verlag, New York, 1981.

[7] J. S. Gibson, *A note on stabilization of infinite dimensional linear oscillations by compact linear feedback*, SIAM J. Control Optim., 18 (1980), pp. 311–316.

[8] A. V. Balakrishnan, *Compensator design for stability enhancement with co-located controllers*, Presented at the IFAC Conference on Control of Distributed Parameters Systems, Perpignan, France, June 1989.

[9] C. E. Benchimol, *A note on the weak stabilizability of contraction semigroups*, SIAM J. Control Optim., 16 (1978), pp. 373–379.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 6, pp. 1300-1321, November 1991

003

FEEDBACK EQUIVALENCE FOR NONLINEAR SYSTEMS AND THE TIME OPTIMAL CONTROL PROBLEM*

B. BONNARD†

Abstract. This article relates the classification of affine control systems under the action of the feedback group, with a differential classification of a set of constrained Hamiltonian vector fields, arising from Pontryagin's Maximum Principle, for the time minimal control problem. They represent the singularities of the input-state mapping. This relation provides a method to compute feedback invariants.

Key words. nonlinear control systems, feedback classification, time optimal control, invariant theory

1. Introduction. Let X, $Y_1, \cdots, Y_m$ be analytic vector fields of $\mathfrak{R}^n$. We are dealing with control systems of the form

$$\frac{dx(t)}{dt} = X(x(t)) + Y(x(t))u(t), \tag{1}$$

where $x \in \mathfrak{R}^n$, $u \in \mathfrak{R}^m$, $Y = (Y_1, \cdots, Y_m)$. They are called *affine systems.*

Let (X, Y) and (X', Y') be two affine control systems. They are called *feedback equivalent* if there exists a C^ω diffeomorphism ψ of $\mathfrak{R}^n$ and a feedback $u = \alpha(x) + \beta(x)u'$, where $\alpha \in \mathscr{C}^\omega(\mathfrak{R}^n, \mathfrak{R}^m)$, $\beta \in \mathscr{C}^\omega(\mathfrak{R}^n, \mathrm{GL}(m, \mathfrak{R}))$ such that
(i) $X' = \psi * X + \psi * Y \cdot \alpha$,
(ii) $Y' = \psi * Y \cdot \beta$,
where $\psi*$ is defined by the following. Let Z be a vector field; then

$$\psi * Z = \frac{\partial \psi^{-1}}{\partial x} \quad (Z \circ \psi)$$

($\psi * Z$ is called the *image* of Z). This action defines a group structure on the set of triplets (ψ, α, β). This group is denoted by G_f.

The classification of linear controllable systems under the action of the Lie subgroup of G_f of triplets (ψ, α, β), where ψ, α are *linear* mappings and β is a *constant* mapping, is now well understood (see [8]). A complete set of (arithmetic) invariants is the set of controllability indices. Canonical forms, called Brunovsky's forms, have been exhibited.

The (local) feedback equivalence problem of system (1) with a linear system was solved around 1980. A chain of distributions is introduced and Brunovsky's canonical forms are recovered by integration of these distributions (see [19] and [18]).

In this article, we study the feedback classification for *generic* nonlinear systems. Our approach is coming from the philosophy of singularity theory and is justified by the following two remarks.

First, consider (ii). Let Δ_Y (respectively, $\Delta_{Y'}$) be the map $x \rightarrow$ the linear span of $\{Y_1(x), \cdots, Y_m(x)\}$ (respectively, $\{Y'_1(x), \cdots, Y'_m(x)\}$. Then, (ii) means that the two distributions Δ_Y and $\Delta_{Y'}$ are diffeomorphic. Therefore, a *first step* in the feedback classification of affine control systems is the *classification of distributions.* The classical results in this area are well known (see [15] for an exposition). They are Frobenius'

* Received by the editors December 7, 1988; accepted for publication (in revised form) July 19, 1990.

† Laboratoire d'Automatique de Grenoble, Unité de Recherche Associée Centre National de la Recherche Scientifique 228-Ecole Nationale Supérieure d'Ingénieurs Electriciens de Grenoble-Institut National Polytechnique de Grenoble, Domaine Universitaire, B.P. 46, 38402 Saint-Martin-d'Hères, France.

theorem for an involutive distribution and Darboux's classification for a generic distribution of codimension one. For recent developments in this theory see [12], but, in any case, all the results are *local* and mainly concern distributions with *constant rank.*

Now, when $Y=0$, (i) shows that the feedback classification is reduced to the analytic classification of vector fields. Near a point x_0 such that $X(x_0)\neq 0$, this problem is solved by the theorem of existence of solutions (X is equivalent to $\partial/\partial x_1$). Near a singular point x_0, i.e., $X(x_0)=0$, this classification is a difficult problem. It was first studied by Lyapunov, Poincaré, and Dulac (see [1]) and is still the object of investigations (see [24] and [26]). If we introduce the distribution $x\to\Re X(x)$, it can be interpreted as a classification problem for distributions of dimension one and with singularities.

The main contribution of this article is to show that with certain (reasonable) assumptions, the feedback classification of pairs (X, Y) is equivalent to the classification of the distribution Δ_Y and a vector field $Z_{\hat{H}}$, which will be defined later. These objects will be encoded in a *constrained vector field.*

Let us make this assertion precise. A system can be viewed as a map, called the *input-state mapping*, which assigns to an input $u(\cdot)$, the response of (1) denoted by $x(\cdot, x_0, u)$, where $x(0)=x_0$ is fixed. If we endow the set of inputs with the L^∞-norm, this map is differentiable and has *singularities*, which are called the *singular trajectories in the time optimal control problem* [5]. They can be parametrized by *Pontryagin's Maximum Principle* (PMP) (see [22]). If we introduce the Hamiltonian H, defined by

$$H(x, p, u)=\langle p, X(x)+Y(x)u\rangle,$$

where $p\in\Re^n\backslash\{0\}$ and $\langle\ ,\ \rangle$ denotes the standard inner product, the singular trajectories are the projections on the x-space of the *constrained Hamiltonian equation*

$$\dot{x}=\frac{\partial H}{\partial p},\qquad \dot{p}=-\frac{\partial H}{\partial x},\qquad (x, p)\in\Sigma,$$

where Σ is the surface $\{(x, p); \langle p, Y(x)\rangle=0\}$.

This equation defines two objects: *a surface* Σ and the solutions $(x(t), p(t))$ of the Hamiltonian equation *that stays in* Σ. We will show that they are the solutions of a vector field $Z_{\hat{H}}$ defined on Σ if m is even or on a proper subset of Σ if m is odd.

We prove that a singularity of the input-state mapping is feedback invariant. Therefore, we can define the action of the feedback group on this set of constrained Hamiltonian equation as follows: $(\psi, \alpha, \beta)\in G_f$ acts on $(\Sigma, Z_{\hat{H}})$ as a symplectic diffeomorphism of $\Re^{2n}$, $\boldsymbol{\psi}$ given by $x=\psi(y)$, $p=q(\partial\psi^{-1}/\partial y)$ (p, q must be written as row vectors). According to Klein, the pair $(\{\Sigma, Z_{\hat{H}}\}, G_f)$ defines a *geometry.*

Now we can formulate the main results of this article. *If two systems are feedback equivalent, then their associated Hamiltonian equations are G_f-equivalent. Moreover, if we restrict our classification to the set of pairs* (X, Y) *such that* (i) Δ_Y *is of constant rank, and* (ii) $m\leqslant n-1$ *when n is even or* $m\leqslant n-2$ *when n is odd, then, in general, the feedback classification problem is equivalent to the classification of our set of constrained Hamiltonian equations.*

"In general" means that our result is valid if we substract to the set of systems a *bad set.* This set is, roughly, the family of systems where the set of singular trajectories is too small to be a complete covariant, i.e., to separate all the orbits for the G_f-action on pairs (X, Y). It is worth pointing out that the *linear systems are in this bad set,* because the input-state mapping of a linear controllable system has no singularity.

The interest of our approach is the following. Understanding a geometry (E, G) consists mainly in describing the algebra of functions $I: E\to\Re$, which are constant on each orbit. They are called the *invariants.*, Now, the invariants of the geometry $(\{\Sigma, Z_{\hat{H}}\}, G_f)$ are feedback invariants. They can be more easily computed than the invariants of the geometry $(\{X, Y\}, G_f)$, because the feedback acts trivially.

Computing invariants is, in general, a dreadful problem. The second part of this article will indicate how to compute feedback invariants by using two nontrivial examples. The first example concerns the feedback invariants arising from the time optimality problem. Indeed, the singular trajectories are the solutions of PMP associated with the *time minimal or time maximal* control problem. Their *optimality status* is a feedback invariant. This problem will be studied when $n=2$ or 3. (When $n>3$, see [6].)

The second example concerns the classification of the vector field $Z_{\hat{H}}$ when we study the feedback classification of a class of homogeneous control systems. We connect this problem with the *linear classification of tensors* (vectors, forms, etc.). All the machinery of this theory can be used, in particular, to compute *polynomial* invariants (see [16] for a good description of this machinery). These computations are outlined for systems in $\mathfrak{R}^3$ and some feedback invariants are connected to the behaviours of the integral curves of $Z_{\hat{H}}$. This example is instructive. It shows that the vector field $Z_{\hat{H}}$ alone can be very rich and *encode informations concerning the optimality problem and the distribution* Δ_Y.

In the conclusion, we indicate how to deal with the feedback classification problem for *nonaffine* control systems. Again, the basic tool is *Pontryagin's Maximum Principle.*

2. Preliminaries.

2.1. Notation and definitions. Let M and N be two analytic manifolds. We denote by $\mathscr{V}^\omega(M)$ the set of analytic vector fields of M. If $M=\mathfrak{R}^n$, a vector field is identified with a map from $\mathfrak{R}^n$ into $\mathfrak{R}^n$. We denote by $\mathscr{C}^\omega(M, N)$ the set of analytic maps from M into N. Let G_d be the group of C^ω-diffeomorphisms of $\mathfrak{R}^n$. The coordinates of $\mathfrak{R}^{2n}=\mathfrak{R}^n\times\mathfrak{R}^n$ are denoted by (x, p). Let $x=(x_1, \cdots, x_n)$, $p=(p_1, \cdots, p_n)$ and endow $\mathfrak{R}^{2n}$ with its canonical symplectic structure defined by $\Omega=\sum_{i=1}^n dx_i\wedge dp_i$. Let Z be a vector field of $\mathfrak{R}^n$, and let g be the map from $\mathfrak{R}^{2n}$ into $\mathfrak{R}$ defined by $(x, p)\rightarrow\langle p, Z(x)\rangle$. The Hamiltonian vector field with Hamilton function g is denoted by $\mathbf{Z}$ and is called the *Hamiltonian lift* of Z.

2.2. Constrained differential equation.

Definition 2.2.1. Let $Z\in\mathscr{V}^\omega(\mathfrak{R}^n)$, $g\in\mathscr{C}^\omega(\mathfrak{R}^n, \mathfrak{R})$, $c\in\mathscr{C}^\omega(\mathfrak{R}^n, \mathfrak{R}^q)$ and set $\tilde{Z}=Z/g$. The pair $(\tilde{Z}, \{c=0\})$ is called a *constrained meromorphic differential equation.* The set $\{c=0\}$ is called the *constraints set.* Let W be the subset of $\mathfrak{R}^n$

$$\{c=0\}\cap\left\{\frac{\partial c}{\partial x}\cdot Z=0\right\},$$

and let us assume that g is not identically zero on W. Let $S=W\cap\{g=0\}$. On $W\backslash S$ there exists an analytic vector field whose solutions are the analytic curves $t\rightarrow x(t)$ that are almost everywhere solutions of $\tilde{Z}$, which stay in $\{c=0\}$.

Definition 2.2.2. Now, we define a *geometry* on the set of meromorphic constrained differential equation. Let G be a subgroup of G_d. With the notation of Definition 2.2.1., two constrained differential equations $(\tilde{Z}, \{c=0\})$ and $(\tilde{Z}', \{c'=0\})$ are said to be *G-equivalent* if there exists $\psi\in G$ such that

(i) $\{c=0\}=\psi(\{c'=0\})$,

(ii) $S=\psi(S')$,

(iii) $\psi*\tilde{Z}|_{W\backslash S}=\tilde{Z}'|_{W'\backslash S'}$.

Observe that since W is the union of S and integral curves of $\tilde{Z}|_{W\backslash S}$, then $W=\psi(W')$.

CONVENTION 2.2.3. The previous geometry will be denoted by $(\{\Sigma, \tilde{Z}|_{W\setminus S}\}, G)$. By solution of a constrained meromorphic differential equation $(\tilde{Z}, \{c=0\})$, we mean an integral curve of $\tilde{Z}|_{W\setminus S}$.

2.3. Definitions. Let E and F be two $\mathfrak{R}$-vector spaces, and let G be a group acting linearly on E and F. A homomorphism $\chi: G \to \mathfrak{R}\setminus\{0\}$ is called a *character*. Let χ be a character. A *semi-invariant* of weight χ is a map $\lambda: E \to \mathfrak{R}$ such that for all $g \in G$, for all $x \in E$, $\lambda(g\cdot x) = \chi(g)\lambda(x)$. It is an *invariant* if $\chi = 1$. A map $\lambda: E \to F$ is a *semi-covariant*, of weight χ, if for all $g \in G$, for all $x \in E$, $\lambda(g\cdot x) = \chi(g) g \cdot \lambda(x)$. It is called a *covariant* if $\chi = 1$.

2.4. Singular trajectories.

DEFINITION 2.4.1. Consider a system of $\mathfrak{R}^n$

$$\frac{dx(t)}{dt} = f(x(t), u(t)), \tag{2}$$

where $u(t)$ is a bounded measurable map from an interval $[0, T]$ into $\mathfrak{R}^m$ and f is a C^1 map from $\mathfrak{R}^n \times \mathfrak{R}^m$ into $\mathfrak{R}^n$.

Let u be a control defined on $[0, T]$ and take $x_0 \in \mathfrak{R}^n$. The response of system (2) to u, initiating from x_0, is an absolutely continuous map $t \to x(t, x_0, u)$ defined on a subinterval $[0, T']$ of $[0, T]$, solution of (2) for almost every $t \in [0, T']$ and such that $x(0) = x_0$. Consider a control u_0 defined on $[0, T]$ such that the response $t \to x(t, u_0, x_0)$ is defined on the whole interval $[0, T]$. Endow the set of controls defined on $[0, T]$ with the L^∞-norm, i.e., $\|u\| = \sup_{t\in[0,T]} |u(t)|$. Let x_0, T be fixed. We denote by $E^{x_0,T}$ the *input-state mapping*, which is given by $u \to x(T, x_0, u)$, and is defined in an open set denoted by $u_0 + V$, where V is a neighborhood of 0. The control u_0 is called *singular* on $[0, T]$ if the Fréchet derivative of $E^{x_0,T}$, denoted by $dE^{x_0,T}$, is not of full rank at u_0. The corresponding response $t \to x(t, x_0, u_0)$ is called a *singular trajectory* (on $[0, T]$).

LEMMA 2.4.2. *Let $v \in V$ and let $A(t) = (\partial f/\partial x)\ (x(t, x_0, u_0), u_0(t))$, $B(t) = (\partial f/\partial u)(x(t, x_0, u_0), u_0(t))$, then*

$$dE^{x_0,T}_{u_0}(v) = M(T) \int_0^T M^{-1}(t) B(t) v(t)\, dt,$$

where $M(t)$ is the $n \times n$ matrix solution of

$$\dot{M}(t) = A(t)M(t),\ M(0) = \text{identity}.$$

Proof. This result is well known (see, for instance, [14]).

PROPOSITION 2.4.3. *Consider, as before, a system of $\mathfrak{R}^n$:*

$$\dot{x}(t) = f(x(t), u(t)).$$

Then the pair u_0 and $x(t, x_0, u_0)$ is singular on $[0, T]$ if and only if there exists a (row) vector $p(t) \in \mathfrak{R}^n\setminus\{0\}$ solution for almost every $t \in [0, T]$ of the adjoint equation

$$\dot{p}(t) = -p(t)A(t), \tag{3}$$

and satisfying for almost every $t \in [0, T]$

$$\langle p(t), B(t)\rangle = 0. \tag{4}$$

Proof. By definition, u_0 is singular on $[0, T]$ if and only if the dimension of the linear span of

$$\left\{\int_0^T M(T)M^{-1}(t)B(t)v(t)\, dt;\quad v \in V\right\}$$

is less than n. Since V is a neighborhood of 0, there exists a (row) vector $\bar{p} \in \mathfrak{R}^n \backslash \{0\}$ such that

$$\bar{p}M(T)M^{-1}(t)B(t)=0 \quad \text{for almost every } t \in [0, T].$$

Let $p(t)=\bar{p}M(T)M^{-1}(t)$; then $p(t)$ is a solution of (3), satisfying (4) for almost every $t \in [0, T]$.

DEFINITION 2.4.4. A pair $(x(t), p(t))$ satisfying (2), (3), and (4) is called a *singular extremal* (corresponding to u_0). Let us denote by Σ the set

$$\left\{(x, p, u) \in \mathfrak{R}^n \times \mathfrak{R}^n \times \mathfrak{R}^m \text{ subject to } \left\langle p, \frac{\partial f}{\partial u}(x, u)\right\rangle = 0\right\}.$$

Let $\tilde{H}$ be the map from $\mathfrak{R}^n \times \mathfrak{R}^n \times \mathfrak{R}^m$ into $\mathfrak{R}$ defined by

$$\tilde{H}(x, p, u)=\langle p, f(x, u)\rangle.$$

The set of equations (2), (3), and (4) can be written as

$$\dot{x}=\frac{\partial \tilde{H}}{\partial p}, (2); \quad \dot{p}=-\frac{\partial \tilde{H}}{\partial x}, (3); \quad \frac{\partial \tilde{H}}{\partial u}=0, (4).$$

According to (2.1), (2) and (3) represent the Hamiltonian lift of system $x \to f(x, \cdot)$.

COROLLARY 2.4.5. *The set of singular extremals contains the solutions of* PMP *for the* time minimal and time maximal *control problem of system* (2).

3. Singular trajectories.

3.1. Computations of singular controls. From now on, we will consider C^ω affine control systems

$$\dot{x}(t)=X(x(t))+Y(x(t))u(t),$$

where $x \in \mathfrak{R}^n$, $u \in \mathfrak{R}^m$, and $Y=(Y_1, \cdots, Y_m)$. The surface Σ defined by $\{(x, p, u); \langle p, Y(x)\rangle=0\}$ is now interpreted as a subset of $\mathfrak{R}^n \times \mathfrak{R}^n$. Consider the constraints (4) written as

$$\text{(5)} \qquad \langle p(t), Y_i(x(t))\rangle=0 \quad \forall i=1, \cdots, m.$$

By Proposition 2.4.3, a singular extremal $(x(t), p(t))$ must satisfy relation (5) for almost every $t \in [0, T]$. Since $t \to (x(t), p(t))$ is continuous, (5) must be true for every $t \in [0, T]$.

For generic pairs (X, Y) (in Whitney's topology), the singular controls will be computed up to a negligible set of singularities as a (dynamic) feedback $(x, p) \to \hat{u}(x, p)$. Its role is to make a maximal subset of Σ invariant for the solutions of (2) and (3). Let us carry out the computations.

Differentiating, with respect to t, the relations

$$\langle p(t), Y_i(x(t))\rangle=0 \quad \forall i=1, \cdots, m$$

and since $x(t)$, $p(t)$ are solutions of (2) and (3), with $f=X+Yu$, we get the equation

$$\text{(6)} \qquad L(x(t), p(t))+O(x(t), p(t))u(t)=0,$$

where $O(x, p)$ is the $m \times m$ matrix (O_{ij}) with $O_{ij}=\langle p, [Y_j, Y_i](x)\rangle$, $L(x, p)$ is the $m \times 1$ matrix (L_i) with $L_i=\langle p, [X, Y_i](x)\rangle$ and where the Lie brackets are computed with the convention

$$[Z_1, Z_2](x)=\frac{\partial Z_2}{\partial x}(x)Z_1(x)-\frac{\partial Z_1}{\partial x}(x)Z_2(x).$$

Let s be the maximal rank of $O(x, p)$. By the antisymmetry property of the Lie bracket, O is an antisymmetric matrix. Let $\mathcal{S}_1$ be the set of pairs (X, Y) such that when m is even, $s = m$, or when m is odd, $s = m-1$. To compute the singular controls, we must distinguish two cases.

Case 1. *The number of inputs is even.* Let (X, Y) be in S_1 and let $\hat{u}$ be defined on an open dense subset of $\mathfrak{R}^n \times \mathfrak{R}^n$ by

$$\hat{u}(x, p) = -O^{-1}(x, p)L(x, p). \tag{7}$$

Now, choose (X, Y) such that det O is not identically zero on Σ. Define the singular control almost everywhere in Σ by $u(t) = \hat{u}(x(t), p(t))$.

Case 2. *The number of inputs is odd.* The computation is similar to the previous one, the only complication being the existence of a kernel for O. Let (X, Y) be in $\mathcal{S}_1$; then the dimension of ker $O(x, p)$ is one for almost every (x, p). Let $r_0 = (x_0, p_0)$ be such a point. Then (6) implies $L(r_0) = 0$ when $u \in \ker O(r_0)$.

This relation defines an additional constraint, which must be satisfied for a singular extremal. This is made precise by the following computation.

Near r_0 (all our computations are local), there exists a matrix $P(r) \in O(r)$, with $r = (x, p)$, such that

$$P^{-1}(r)O(r)P(r) = (\bar{O}_1, 0),$$

where $\bar{O}_1(r)$ is an antisymmetric $(m-1) \times (m-1)$ matrix. Moreover, $P(r)$ can be chosen analytic in the coefficients of $O(r)$, since the eigendirection corresponding to the zero eigenvalue of O depends analytically on the coefficients of O.

$P^{-1}L$ can be written as $(\bar{L}_1, \bar{L}_2)$ where $\bar{L}_1$ is an $(m-1)$ column vector, and $\bar{L}_2$ is a scalar. Let us write $\bar{u} = P^{-1}u$ as $\bar{u}_1 + \bar{u}_2$, where $\bar{u}_1 \in \mathfrak{R}^{m-1}$, $\bar{u}_2 \in \mathfrak{R}$. Then (6) is equivalent to

$$\bar{L}_1(r) + \bar{O}_1(r)\bar{u}_1 = 0, \tag{6a}$$

$$\bar{L}_2(r) = 0. \tag{6b}$$

Let $\hat{\bar{u}}_1$ be the map defined almost everywhere on $\mathfrak{R}^{2n}$ by

$$\hat{\bar{u}}_1(r) = -\bar{O}_1^{-1}(r)\bar{L}_1(r). \tag{7a}$$

To compute the component of the singular control in the kernel of O, we derive (6b) along a solution of (2) and (3). Observe that $\bar{L}_2(r)$ is near r_0, analytic in the coefficients of $O(r)$ and $L(r)$. They are of the form $\langle p, Z(x)\rangle$, where Z belongs to the Lie algebra generated by $\{X, Y_1, \cdots, Y_m\}$. Therefore, differentiating relation (6b) with respect to t along a solution $r(t)$ of (2), (3), we get a relation of the form

$$f_1(r(t)) + g_1(r(t))\bar{u}_1(t) + \bar{u}_2(t)g_2(r(t)) = 0. \tag{6c}$$

Let (X, Y) be chosen such that g_2 is not identically zero and let $\hat{\bar{u}}_2$ be the map defined almost everywhere on $\mathfrak{R}^{2n}$ by

$$\hat{\bar{u}}_2(r) = -(f_1(r) - g_1(r)\hat{\bar{u}}_1(r))g_2^{-1}(r). \tag{7b}$$

A control $u(t) = P\bar{u}(t)$ defined by the analytic restriction of $(\hat{\bar{u}}_1(r(t)), \hat{\bar{u}}_2(r(t))$ to $\Sigma \cap \{\bar{L}_2(r) = 0]$ is a singular control.

3.2. Notation. Let M be the variety defined by (i) when m is even, $\Sigma = M$, and (ii) when m is odd $M = \Sigma \cap \{\bar{L}_2 = 0\}$. Observe that in each case the number of equations defining M is *even.*

Let $\mathcal{S}_g$ (g means "good") be the set of pairs (X, Y) in $\mathcal{S}_1$ given by (i) when m is even, det O is not identically 0 on Σ, and (ii) when m is odd, O is almost everywhere on M of rank $m-1$ and g_2 is almost everywhere invertible on M.

Take $(X, Y) \in \mathcal{S}_g$. Let $\hat{u}$ be the map almost everywhere defined from $\mathfrak{R}^n \times \mathfrak{R}^n$ into $\mathfrak{R}^m$ (by (i) when m is even, $\hat{u}$ is given by (7), and (ii) when m is odd, $\hat{u} = P\hat{\bar{u}}$ is given by (7a) and (7b)). Let S be the of points of M such that $\hat{u}$ is not analytic.

From 3.1, on $M\backslash S$, the singular control is uniquely defined by the analytic map $\hat{u}$. In particular, in $M\backslash S$, every singular extremal corresponds to a unique singular control, whose role is to make $M\backslash S$ invariant for the solutions of (2) and (3). Such a singular extremal is said to be of *minimal order.* Let us denote by $\hat{H}$ the function defined almost everywhere on $\mathfrak{R}^{2n}$ by

$$\hat{H}(x, p) = \langle p, X(x) + Y(x)\hat{u}(x, p)\rangle$$

and let $\tilde{Z}_{\hat{H}}$ be the associated meromorphic Hamiltonian vector field. Let us denote $Z_{\hat{H}}$ its restriction to $M\backslash S$.

PROPOSITION 3.3. *Let (X, Y) be a system of $\mathcal{S}_g$. Then the singular extremals of* minimal order *are the solutions of the constrained Hamiltonian differential equation*

$$\dot{x} = \frac{\partial \hat{H}}{\partial p}, \qquad \dot{p} = -\frac{\partial \hat{H}}{\partial x}, \qquad (x, p) \in \Sigma. \tag{8}$$

Proof. By using the constraints $\langle p, Y(x)\rangle = 0$, we get

$$\frac{\partial \hat{H}}{\partial p} = X + Y\hat{u} + \left\langle p, Y\frac{\partial \hat{u}}{\partial p}\right\rangle = X + Y\hat{u} \quad \text{on } \Sigma$$

and

$$\frac{\partial \hat{H}}{\partial x} = \left\langle p, \frac{\partial X}{\partial x} + \frac{\partial Y}{\partial x}\hat{u}\right\rangle + \left\langle p, Y\frac{\partial \hat{u}}{\partial x}\right\rangle$$

$$= \left\langle p, \frac{\partial X}{\partial x} + \frac{\partial Y}{\partial x}\hat{u}\right\rangle \quad \text{on } \Sigma.$$

Remarks 3.4. On $M\backslash S$, a singular control can be interpreted as an unique *static* feedback such that Σ is made invariant for the solutions of the lifted system $\mathbf{X} + \mathbf{Y}u$.

The singular controls have been computed for a generic system. In the nongeneric case, the computation is similar. For instance, if the distribution Δ_Y is involutive, then O is identically zero on Σ, since, in this case, $\langle p, Y_i(x)\rangle = 0$ for all i implies $\langle p, [Y_j, Y_i]$-$(x)\rangle = 0$. Thus, the equation $L(x(t), p(t)) = 0$ must be satisfied and can be differentiated with respect to t to compute the singular controls (see § 6 for more details).

Only the singular extremals of minimal order have been computed. Computations outlined in [4] indicate that for generic pairs (X, Y), they are the only singular extremals. Moreover, if they exist they are contained in S, and [2] shows heuristically that they can be analyzed as solutions of the *meromorphic* differential equations associated to $\tilde{Z}_{\hat{H}}$ restricted to M.

4. Feedback classification and singular trajectories

DEFINITION 4.1. Let λ be the map that associates to a system (X, Y) of $\mathcal{S}_g$ the constrained differential equation (8); i.e., λ is the map $(X, Y) \to (\Sigma, Z_{\hat{H}})$.

Let $(\psi, \alpha, \beta) \in G_f$ and we lift ψ into a symplectic diffeomorphism $\boldsymbol{\psi}$ of $\mathfrak{R}^n \times \mathfrak{R}^n$ defined by

$$x = \psi(y), \qquad p = q\frac{\partial \psi^{-1}}{\partial y},$$

where p and q are row vectors. The action of (ψ, α, β) on $\lambda(X, Y)$ is by definition reduced to the action of $\boldsymbol{\psi}$ on $(\Sigma, Z_{\hat{H}})$ as given in Definition 2.2.2.

In other words, the feedback acts trivially and ψ acts as the symplectic change of coordinates $\boldsymbol{\psi}$ on $(\Sigma, Z_{\hat{H}})$. The corresponding geometry is denoted by $(\{\Sigma, Z_{\hat{H}}\}, G_f)$.

THEOREM 4.2. *The following diagram is commutative*:

$$\begin{tikzcd}
\mathcal{S}_g \arrow[r, "\lambda"] \arrow[d, "G_f"'] & \lambda(\mathcal{S}_g) \arrow[d, "G_f"] \\
\mathcal{S}_g \arrow[r, "\lambda"] & \lambda(\mathcal{S}_g)
\end{tikzcd}.$$

In other words, λ is a covariant.

Proof. Every $g \in G_f$ can be written as $g_2 \cdot g_1$ with $g_1 = (\psi, 0, \mathrm{id})$, $g_2 = (\mathrm{id}, \alpha, \beta)$. Take $(X, Y) \in \mathcal{S}_g$ and set $(X', Y') = g_1 \cdot (X, Y), (X'', Y'') = g_2 \cdot (X', Y')$. Let $\hat{H}$ (respectively, $\hat{H}'$, $\hat{H}''$), Σ (respectively, Σ', Σ''), M (respectively, M', M''), $\hat{u}$ (respectively, $\hat{u}'$, $\hat{u}''$), $Z_{\hat{H}}$ (respectively, $Z'_{H'}$, $Z''_{\hat{H}''}$), and S (respectively, S', S'') be the elements defined in § 3.2 and associated to system (X, Y) (respectively, $(\mathrm{X}', \mathrm{Y}')$, $(\mathrm{X}'', \mathrm{Y}'')$).

First, let us study the action of g_1, i.e., the action of the symplectic diffeomorphism $\boldsymbol{\psi}$ defined by

$$x = \psi(y), \qquad p = q\frac{\partial \psi^{-1}}{\partial y},$$

on those objects. Observe that the constraints $\langle p, Y(x)\rangle = 0$ are written in the (y, q) coordinates as $\langle q(\partial\psi^{-1}/\partial y),\ Y(\psi(y))\rangle = 0$. Since q is a row vector, this is equivalent to $\langle q, (\partial\psi^{-1}/\partial y)Y(\psi(y))\rangle = 0$, i.e., $\langle q, \psi * Y(y)\rangle = 0$. Hence $\boldsymbol{\psi}$ maps Σ' into Σ. Similarly, from § 3.1, we have that $\boldsymbol{\psi}$ maps M' onto M, S' onto S, and $\hat{u}' = \hat{u} \circ \boldsymbol{\psi}$ almost everywhere on $\mathfrak{R}^{2n}$. Thus we have

$$\begin{aligned}\hat{H}(x, p) &= \langle p, X(x) + Y(x)\hat{u}(x, p)\rangle = \left\langle q\frac{\partial\psi^{-1}}{\partial y}, X(\psi(y)) + Y(\psi(y))\hat{u}(\boldsymbol{\psi}(y, q))\right\rangle \\ &= \langle q, (\psi * X)(y) + (\psi * Y)(y)\hat{u}'(y, q)\rangle = \hat{H}'(y, q).\end{aligned}$$

Therefore the symplectic transformation $\boldsymbol{\psi}$ maps the solutions of the Hamiltonian equation of $\mathfrak{R}^{2n}$ defined by $\hat{H}$ onto those defined by $\hat{H}'$. In particular, $\psi * Z_{\hat{H}} = Z_{\hat{H}'}$.

Now let us study the action of the feedback $u' = \alpha(y) + \beta(y)u''$. Since $\beta(y)$ is invertible, this transformation induces a bijection on the set of inputs and maps biunivoquely the singular controls associated to (X', Y') onto those associated to (X'', Y''). Moreover, $(y(t), q(t))$ is a singular extremal associated to u' if and only if it is singular extremal associated to u''. This is geometrically obvious and can be proved as follows.

First, observe that $\Sigma' = \Sigma''$ because the constraints $\langle q, Y'(y)\rangle = 0$ and $\langle q, (Y'\beta)(y)\rangle = 0$ are equivalent, since $\beta(y) \in \mathrm{GL}(m, \mathfrak{R})$. Now, by using (2), (3), and (4), observe that if $(y(t), q(t))$ is a singular extremal corresponding to $u'(t)$ for (X', Y'), then it is a singular extremal corresponding to $u''(t)$ for (X'', Y'') and reciprocally.

In particular, $M' = M''$ almost everywhere since they are almost everywhere the union of singular extremals, and by continuity $M' = M''$. Moreover, by definition, $u'(t) = \alpha(y(t)) + \beta(y(t))u''(t)$, then $\hat{u}'(y, q) = \alpha(y) + \beta(y)\hat{u}''(y, q)$ on M'. Hence $S' = S''$. Therefore $\hat{H}'(y, q) = \langle q, X'(y) + Y'(y)\hat{u}'(y, q)\rangle = \langle q, X''(y) + Y''(y)\hat{u}''(y, q)\rangle = \hat{H}''(y, q)$, on $M' \backslash S'$. Then $Z_{\hat{H}'} = Z_{\hat{H}''}$. (The results can be obtained by direct computations.) □

For further studies, it is interesting to point out more geometry connected with the singular problem.

Remark 4.3. To generalize our theory to the C^∞ case, we remark that we may deal with a differential equation defined on the whole set M. This follows by taking into account the invariant meaning of the set S. For simplicitly, let us make this assertion more precise in the single input case. The variety M is then defined by $H_1=0$, $H_2=0$, where $H_1=\langle p, Y(x)\rangle$ and $H_2=\langle p, [X, Y](x)\rangle$. The set S is then defined by $H_1=H_2=\{H_1, H_2\}=0$, where $\{\ ,\ \}$ is the Poisson bracket. The singular control is given on $M\backslash S$ by

$$\hat{u}(x,p)=-\frac{\langle p,[X,[X, Y]](x)\rangle}{\langle p,[Y,[X, Y]](x)\rangle}.$$

Let us introduce the maps $f_1^{X,Y}:(x,p)\to\langle p,[X,[X, Y]](x)\rangle$ and $f_2^{X,Y}:(x,p)\to\langle p,[Y,[X, Y]](x)\rangle$. Now we analyze the action induced by G_f on the f_i's.

Let ψ be a C^ω-diffeomorphism: $x=\psi(y)$ and let $\boldsymbol{\psi}$ defined as usual by: $x=\psi(y)$, $p=q(\partial\psi^{-1}/\partial y)$. Set $X'=\psi * X$, $Y'=\psi * Y$.

Then we have

$$\begin{aligned}f_1^{X,Y}(x,p)&=\langle p,[X,[X, Y]](x)\rangle=\left\langle q\frac{\partial\psi^{-1}}{\partial y},[X,[X, Y]](\psi(y))\right\rangle\\&=\langle q,\psi*[X,[X, Y]](y)\rangle=<q,[\psi*X,[\psi*X,[\psi*Y]](y)\rangle\\&=f_1^{X',Y'}(\boldsymbol{\psi}^{-1}(x,p))\end{aligned}$$

and, similarly, $f_2^{X,Y}=f_2^{X',Y'}\circ\boldsymbol{\psi}^{-1}$.

Now let u be the feedback $\alpha(x)+\beta(x)u'$ and let $X'=(\mathrm{id},\alpha,\beta)\cdot X$, $Y'=(\mathrm{id},\alpha,\beta)\cdot Y$. Then, *on M* we have

$$f_1^{X',Y'}=\beta(f_1^{X,Y}+\alpha f_2^{X,Y}),\qquad f_2^{X',Y'}=\beta^2 f_2^{X,Y}.$$

Therefore, if we replace (8) by the *analytic* differential equation

$$(8')\qquad \frac{dx}{d\tau}=f_2\frac{\partial\hat{H}}{\partial p},\qquad\frac{dp}{d\tau}=-f_2\frac{\partial\hat{H}}{\partial x},\qquad(x,p)\in\Sigma,$$

and making G_f acting on f_2 and (8′) by *change of coordinates* $\boldsymbol{\psi}$ *only*, the previous results tell us that the maps

(i) $(X, Y)\to f_2^{X,Y}$ restricted to M,

(ii) $(X, Y)\in\mathcal{S}_g\to f_2\cdot\tilde{Z}_{\hat{H}}$ restricted to M ($\tilde{Z}_{\hat{H}}$ is defined in (3.2))

are *semicovariants.*

This has the following geometric interpretation. S is $\{f_2=0\}\cap M$, and the solutions $(x(\tau),p(\tau))$ of (8′) in $M\backslash S$ are the singular extremals $(x(t),p(t))$ of minimal order, reparametrized by the transformation $f_2\,dt=d\tau$. With this trick, we get an analytic vector field, defined on the whole M.

Remark 4.4. Now we will show that $Z_{\hat{H}}$ can be interpreted as an Hamiltonian vector field defined almost everywhere on M. for simplicity, let us investigate the single input case.

The variety M is then given by $\langle p, Y(x)\rangle=\langle p,[X, Y](x)\rangle=0$, and S is the set of points (x,p) of M defined by $\langle p,[Y,[X, Y]](x)\rangle=0$. Let K be the set of points of $\mathfrak{R}^n$ such that Y and $[X, Y]$ are colinear. Observe that K is feedback invariant and thus has an invariant meaning for the G_f action on pairs (X, Y). Assume that (X, Y) is chosen in $\mathcal{S}_g$ such that K is a proper subset of $\mathfrak{R}^n$, and let W be $M\cap(\mathfrak{R}^n\backslash K\times\mathfrak{R}^n)$.

Thus, W is a regular submanifold of $\mathfrak{R}^n \times \mathfrak{R}^n$, with even dimension. Let Ω' be the restriction of the symplectic form $\Omega = dx \wedge dp$ to $W \backslash S$.

Let us assume that $Y = \partial/(\partial x_1)$.

We claim that $(W\backslash S, \Omega')$ is a symplectic manifold. This can be (naively) proved as follows. Recall that x_i and p_i denote the coordinates of x and p, and let us write x and p as $(x_1, \bar{x})$ and $(p_1, \bar{p})$. X can be written as $X_1(\partial/(\partial x_1)) + \bar{X}(\partial/(\partial \bar{x}))$. Since $Y = \partial/(\partial x_1)$, $\langle p, Y(x)\rangle = 0$ is equivalent to $p_1 = 0$ and then $\langle p, [X, Y](x)\rangle = 0$ is equivalent to $\langle \bar{p}, \partial \bar{X}/(\partial x_1)\rangle = 0$. Moreover, if $\langle \bar{p}, (\partial^2 \bar{X})/(\partial x_1^2)\rangle \neq 0$, then by the implicit function theorem, the previous equation can be locally solved as $x_1 = f(\bar{x}, \bar{p})$. Now observe that on W, the condition $\langle \bar{p}, (\partial^2 \bar{X})/(\partial x_1^2)\rangle \neq 0$ is equivalent $\langle p, [Y, [X, Y]](x)\rangle \neq 0$. Moreover, $\Omega' = dx_1 \wedge dp_1 + d\bar{x} \wedge d\bar{p} = df(\bar{x}, \bar{p}) \wedge dp_1 + d\bar{x} \wedge d\bar{p} = d\bar{x} \wedge d\bar{p}$.

Since $Z_{\hat{H}}$ is tangent to $M\backslash S$, *the restriction of* $Z_{\hat{H}}$ *to* $W\backslash S$ *is a Hamiltonian vector field for* $(W\backslash S, \Omega')$ *whose Hamiltonian is the restriction of* $\hat{H}$ *to* $W\backslash S$. This vector field is analytic since $\langle p, [Y, [X, Y]](x)\rangle \neq 0$ on $W\backslash S$. □

Now we study when the feedback classification is equivalent to the G_f-classification of the constrained Hamiltonian equation (8). The first step is to reduce the feedback classification of pairs (X, Y) to the classification of distributions Δ_Y and the "orthogonal complement of X with respect to Δ_Y." This is a useful reduction to compute normal forms for the G_f-action on pairs (X, Y) (see [3] for such computations).

DEFINITION 4.5. We denote by G_d^Y the subgroup of analytic diffeomorphisms of $\mathfrak{R}^n$, leaving invariant the distribution Δ_Y, i.e.,

$$G_d^Y = \{\psi \text{ s.t. } \psi * Y_i \in \Delta_Y, \qquad \forall i = 1, \cdots, m\}.$$

Two vector fields X, X' are called *equivalent modulo* Δ_Y if there exists $\psi \in G_d^Y$ such that $\psi * X = X' \pmod{\Delta_Y}$, i.e., for all $x \in \mathfrak{R}^n$, $(\psi * X - X')(x) \in \Delta_Y(x)$.

LEMMA 4.6. *The following assertions are equivalent*:

(i) (X, Y) *and* (X', Y') *are feedback equivalent*;

(ii) (a) *There exists a* C^ω*-diffeomorphism* ψ_1 *such that* $\psi_1 * \Delta_Y = \Delta_{Y'}$,

(b) $\psi_1 * X$ *and* X' *are equivalent modulo* $\Delta_{Y'}$.

Proof. The proof is straightforward. Assume that (X, Y) and (X', Y') are feedback equivalent. Then there exists $(\psi_1, 0, \beta_1) \in G_f$ such that $(\psi_1, 0, \beta_1) \cdot (X, Y) = (\bar{X}, Y')$, where $\bar{X} = \psi_1 * X$. In other words, $\psi_1 * \Delta_Y = \Delta_{Y'}$.

Now the crucial fact is that we may restrict our analysis to the feedback classification problem, where $\Delta_{Y'}$ is fixed, i.e., for the subgroup action $\{(\psi, \alpha, \beta), \psi \in G_d^{Y'}\}$. Indeed, since $(\bar{X}, Y')$ is feedback equivalent to (X', Y'), then there exists $(\psi_2, \alpha_2, \beta_2) \in G_f$ such that

$$(\psi_2, \alpha_2, \beta_2) \cdot \bar{X} = X' \quad \text{and} \quad \psi_2 * \Delta_{Y'} = \Delta_{Y'}.$$

Hence $\psi_2 \in G_d^{Y'}$.

Now since $\psi_2 \in G_d^{Y'}$, $(\psi_2, \alpha_2, \beta_2) \cdot X$ can be written as $\psi_2 * \bar{X} (\bmod \Delta_{Y'}) + \theta$, where $\theta \in \Delta_{Y'}$. Hence (i)$\Rightarrow$(ii) and, similarly, (ii)$\Rightarrow$(i). □

Assumptions 4.7. Let $\mathcal{S}_r$ be the set of pairs (X, Y) such that

(i) $(X, Y) \in \mathcal{S}_g$,

(ii) The rank of the matrix $Y(x)$ is m, for all $x \in \mathfrak{R}^n$,

(iii) $\pi(M\backslash S)$ contains a (nonempty) open subset of $\mathfrak{R}^n$, where π is the projection $(x, p) \to x$.

THEOREM 4.8. *Let* (X, Y) *and* (X', Y') *be systems of* $\mathcal{S}_r$. *Then the following assertions are equivalent*:

(i) (X, Y) *and* (X', Y') *are feedback equivalent*,

(ii) *The constrained Hamiltonian vector fields $\lambda(X, Y)$ and $\lambda(X', Y')$ are G_f-equivalent.*

Proof. First, we need the following result.

LEMMA. *Let Σ and Σ' denote the constraints sets of (X, Y) and (X', Y'). Then Σ and Σ' are G_f-diffeomorphic if and only if Δ_Y and $\Delta_{Y'}$ are diffeomorphic.*

Proof. Let us assume the existence of a symplectic diffeomorphism ψ defined by $x=\psi(y)$, $p=q(\partial\psi^{-1}/\partial y)$, which maps Σ' onto Σ. Then we have

$$\begin{aligned}(y, q)\in\Sigma' &\Leftrightarrow\langle q, Y'(y)\rangle=0\Leftrightarrow(x, p)\in\Sigma\\ &\Leftrightarrow\langle p, Y(x)\rangle=0\Leftrightarrow\left\langle q\frac{\partial\psi^{-1}}{\partial y}, Y(\psi(y))\right\rangle=0\\ &\Leftrightarrow\langle q, \psi * Y(y)\rangle=0.\end{aligned}$$

Now fix $y\in\Re^n$ and observe that $\langle q, Y'(y)\rangle=0$ coincides with $\langle q, \psi * Y(y)\rangle=0$ if and only if there exists a unique $\beta(y)\in\mathrm{GL}\,(m, \Re)$ such that $Y'(y)\beta(y)=(\psi * Y)(y)$, since the rank of Y and Y' is m. Moreover, as $y\to Y(y)$, $Y'(y)$ are analytic, then $y\to\beta(y)$ is analytic.

We proved that ψ maps Σ' onto Σ if and only if $\Delta_{Y'}=\Delta_{\psi * Y}$, i.e., $\Delta_{Y'}$ and Δ_Y are ψ-diffeomorphic. □

Now we proceed to the proof of Theorem 4.8. From Theorem 4.2, (i)⟹(ii), and it remains to prove (ii)⟹(i).

Assume that the constrained vector field $\lambda(X, Y)$ and $\lambda(X', Y')$ are G_f-equivalent. Thus the varieties Σ and Σ' are G_f-diffeomorphic and then we may assume $\Sigma=\Sigma'$. By the previous lemma, it means $\Delta_Y=\Delta_{Y'}$.

Now since $\lambda(X, Y)$ and $\lambda(X', Y')$ are G_f-equivalent and $\Sigma=\Sigma'$, then there exists a diffeomorphism ψ of $\Re^n$ such that ψ preserves $\Sigma=\Sigma'$ and satisfies

$$\psi * Z_{\hat H}=Z_{\hat H'}\quad\text{on } M'\backslash S'.$$

Therefore $\psi\in G_d^Y$ and we have on $M'\backslash S'$ the following:

$$\begin{aligned}\psi * Z_{\hat H}&=\psi *\left(\frac{\partial\hat H}{\partial p}\frac{\partial}{\partial x}-\frac{\partial\hat H}{\partial x}\frac{\partial}{\partial p}\right)\\ &=\left(\psi * X+\psi * Y\cdot u\circ\psi\frac{\partial}{\partial y}+\cdots\frac{\partial}{\partial q}\right)\\ &=Z_{\hat H'}=\left(X'+Y'\hat u'\frac{\partial}{\partial y}+\cdots\frac{\partial}{\partial q}\right).\end{aligned}$$

In particular, on $M'\backslash S'$ we have

$$\psi * X+\psi * Y\cdot\hat u\circ\psi=X'+Y'\hat u.$$

Now since $\psi\in G_d^Y$, we have $\Delta_{\psi * Y}=\Delta_Y=\Delta_{Y'}$ and the previous equation implies that

$$\psi * X=X'(\mathrm{mod}\,\Delta_Y)\quad\text{on } \pi(M'\backslash S').$$

Now, by assumption, this equation holds on a nonempty open subset of $\Re^n$. By analycity, it is true everywhere. Thus, by Lemma 4.6, (ii)⟹(i) and Theorem 4.8 is proved.

COROLLARY 4.9. *Let us assume that n is even, $m\leq n-1$ or n is odd, $m\leq n-2$. Then for an open dense set of pairs (X, Y) (for Whitney's topology) the feedback classification is equivalent to the G_f-classification.*

Proof. M is defined by m equations in the even case or $m+1$ equations in the odd case. They are linear in p, and to eliminate p, we generically need at least $n-1$ equations. Hence, $\pi(M)$ is (generically) a proper subset of $\mathfrak{R}^n$ if n is even, $m > n-1$, or n is odd, $m > n-2$. This proves the result.

5. Optimality and feedback classification. By using two examples, we will show the connection between the feedback classification and the time optimality status of the singular trajectories.

5.1. Systems in $\mathfrak{R}^2$. Consider the following system of $\mathfrak{R}^2$:

$$\dot{v} = X(v) + uY(v), \tag{9}$$

where $v=(x,y)\in\mathfrak{R}^2$. Then, by using the equations $\langle p, Y(v)\rangle = \langle p, [X, Y](v)\rangle = 0$, where $p \in \mathfrak{R}^2\backslash\{0\}$, we see that the singular trajectories are contained in $L=\{v\in\mathfrak{R}^2$ subject to $\det(Y(v), [X, Y](v))\}=0$.

Hence, for an open dense set of systems, Assumption 4.7 (iii) is not satisfied and Theorem 4.8 cannot be applied to reduce the feedback classification of systems (X, Y) to the G_f-classification of pairs $(\Sigma, Z_{\hat{H}})$. This can be illustrated by the following example. Consider the two systems:

$$\text{(A)}\quad \dot{x}=x^2-y^2,\quad \dot{y}=u \qquad \text{(B)}\quad \dot{x}=x^2+y^2,\quad \dot{y}=u.$$

By a straightforward computation, we get that for *both systems*, L is the line $y=0$ on which the singular trajectories are the solutions of $\dot{x}=x^2$. Moreover, the constraints set Σ are the same. Nevertheless, (A) and (B) are not feedback equivalent since, clearly, the singular trajectory on $\{x>0\}$ is *time minimal* for (A) and *time maximal* for (B). Hence, Theorem 4.8, without Assumption 4.7(iii), is not valid. *Since the optimality status of a singular trajectory is not encoded in* $(\Sigma, Z_{\hat{H}})$, *an additional covariant has to be used to separate the orbits of* G_f *acting on systems* (X, Y). To construct this covariant we can proceed as follows.

Let γ be a singular trajectory defined on $[0, T]$ and let us assume that X and Y are never colinear on γ. Following [17], to evaluate the optimality status of γ, we introduce the form $\alpha = r\,dv$ defined in a neighborhood of γ by $\langle r, X\rangle = 1$, $\langle r, Y\rangle = 0$.

This form has the following nice properties:

(i) if γ_1 is a solution of (9) defined on $[0, t_1]$, then $\int_{\gamma_1}\alpha = \int_0^{t_1} dt = t_1$,

(ii) $d\alpha = 0$ on the singular trajectory γ.

Hence, let γ_1 be a solution of (9) starting at $t=0$ from $\gamma(0)$ and with endpoint $\gamma_1(t_1)=\gamma(T)$. Then we have

$$\begin{aligned}\int_\gamma \alpha - \int_{\gamma_1}\alpha &= \int_{\gamma\cup -\gamma_1} d\alpha \quad \text{(Stoke's theorem)}\\ &= T - t_1.\end{aligned}$$

Hence, the optimality status of γ can be characterized. For instance, for system (A), we have $d\alpha<0$ if $y>0$ and $d\alpha>0$ if $y<0$. This shows that the singular trajectory γ on $\{x>0\}$ is time minimal with respect to every solution of (A), in a C^0-neighborhood of γ.

5.2. Systems in $\mathfrak{R}^3$. Consider the following system:

$$\dot{v} = X(v) + uY(v), \tag{10}$$

where $v=(x,y,z)\in\mathfrak{R}^3$. Let

$$D_1^{X,Y} = \det(Y, [X, Y], [Y, [X, Y]]), \qquad D_2^{X,Y} = \det(Y, [X, Y], [X, [X, Y]]),$$

and $D_3^{X,Y} = \det(Y, [X, Y], X)$. (They will be denoted by D_i when no confusion is possible.)

Convention 5.2.1. For the remainder of this section, we consider only systems (X, Y) such that $D_1^{X,Y}$ is not identically zero. By singular trajectory of (X, Y), We mean a singular trajectory contained in $\mathfrak{R}^3 \backslash \{D_1^{X,Y} = 0\}$.

LEMMA 5.2.2. *The singular trajectories of* (X, Y) *are the solutions of*

$$\dot{v} = X(v) - \frac{D_2(v)}{D_1(v)} Y(v) \tag{11}$$

in $\mathfrak{R}^3 \backslash \{D_1 = 0\}$.

Proof. Since $p \neq 0$, the relations

$$\langle p, Y\rangle = \langle p, [X, Y]\rangle = \langle p, [X, [X, Y]]\rangle + u\langle p, [Y, [X, Y]]\rangle = 0$$

defining the singular control (see § 3.1) imply $D_2 + uD_1 = 0$. Thus, on $\mathfrak{R}^3 \backslash \{D_1 = 0\}$, the singular control is defined by a feedback $\hat{u}(v) = -D_2(v)/D_1(v)$.

DEFINITION 5.2.3. Let X_Y be the vector field on $\mathfrak{R}^3 \backslash \{D_1 = 0\}$ associated with (11), and let λ_π be the map $(X, Y) \to X_Y$. Observe that X_Y is the restriction of $\pi * Z_{\hat{H}}$ to $\mathfrak{R}^3 \backslash \{D_1 = 0\}$ and $\lambda_\pi = \pi \circ \lambda$, where π designs the map $(x, p) \to x$.

The action of G_f on $Z_{\hat{H}}$ induces the following action on X_Y:

$$(\psi, \alpha, \beta) \cdot X_Y = \psi * X_Y,$$

and from Theorems 4.2 and 4.8 we have the following proposition.

PROPOSITION 5.2.4. *The map* λ_π *is a covariant. Moreover, the feedback classification of pairs* (X, Y) *is equivalent to the* G_f*-classification of pairs* (Δ_Y, X_Y).

In particular, the time optimality status of singular trajectories is encoded in (Δ_Y, X_Y). We will briefly describe how this occurs.

LEMMA 5.2.5. *Let* $(\psi, \alpha, \beta) \in G_f$, *then*

(i) $D_i^{\psi * X, \psi * Y}(v) = \det(\partial\psi^{-1}/\partial v) D_i^{X,Y}(\psi(v))$, *for all* $i = 1, 2, 3$;

(ii) $D_1^{X + Y\alpha, Y\beta} = \beta^4 D_1^{X,Y}$;

(iii) $D_2^{X + Y\alpha, Y\beta} = \beta^3(D_2^{X,Y} + \alpha D_1^{X,Y})$;

(iv) $D_3^{X + Y\alpha, Y\beta} = \beta^2 D_3^{X,Y}$.

Proof. These relations can be obtained by straightforward computations. For instance,

$$\begin{aligned} D_1^{\psi * X, \psi * Y}(v) &= \det(\psi * Y(v), [\psi * X, \psi * Y](v), [\psi * Y, [\psi * X, [\psi * Y]]](v)) \\ &= \det(\psi * Y(v), \psi * [X, Y](v), \psi * [Y, [X, Y]](v)) \\ &= \det\left(\frac{\partial \psi^{-1}}{\partial v}\right) \det(Y, [X, Y], [Y, [X, Y]])(\psi(v)), \text{etc.} \end{aligned}$$

COROLLARY 5.2.6. *Let* $f: \mathfrak{R}^3 \to \mathfrak{R}$ *and define the action of* $(\psi, \alpha, \beta) \in G_f$ *on* f *as follows*: $(\psi, \alpha, \beta) \cdot f = f \circ \psi$. *then, we have the following* semicovariants:

(i) $\lambda_1: (X, Y) \to D_1^{X,Y}$,

(ii) $\lambda_3: (X, Y) \to D_3^{X,Y}$.

PROPOSITION 5.2.7. *The sets defined by* $D_3 = 0$, $D_1D_3 > 0$, *and* $D_1D_3 < 0$, *are invariant sets for the solutions of* $\dot{v} = X_Y(v)$.

Proof. On $\mathfrak{R}^3 \backslash \{D_1 = 0\}$, Y and $[X, Y]$ are independent and $D_3 = 0$ is the set of points where X belongs to the linear span of $\{X, [X, Y]\}$. Thus $\langle p, Y\rangle = \langle p, [X, Y]\rangle = 0$ implies $\langle p, X\rangle = 0$ on $D_3 = 0$. Moreover, if γ is a singular trajectory, with adjoint vector p, then the derivative of $t \to \langle p(t), X(\gamma(t))\rangle$ is $\langle p(t), [X, Y](\gamma(t))\rangle$ and is equal to 0.

Thus $X(\gamma(t))$ belongs to the linear span of $Y(\gamma(t))$ and $[X, Y](\gamma(t))$. (A shorter proof is to remark that $D_3=0$ corresponds to singular trajectories with Hamiltonian $H=0$.) Now $D_1D_3>0$ and $D_1D_3<0$ are invariant, since along a singular trajectory, the Lie bracket $[Y,[X, Y]]$ cannot cross the linear span of Y and $[X, Y]$ (by Convention 5.2.1, a singular trajectory is contained in $\mathfrak{R}^3\backslash\{D_1=0\}$).

DEFINITION 5.2.8. A singular trajectory γ is called

(i) exceptional if γ belongs to $D_3=0$,

(ii) hyperbolic if γ belongs to $D_1D_3>0$,

(iii) elliptic if γ belongs to $D_1D_3<0$.

PROPOSITION 5.2.9. *Let γ be a singular trajectory. Then there exists a C^0-neighborhood of γ such that γ is time minimal if γ is exceptional or hyperbolic and time maximal if γ is elliptic, with respect to all solutions of* (10) *contained in the neighborhood of γ if there exists no conjugate point along γ.*

The concept of conjugate point and the proof of this result are given in [6].

Remark 5.2.10. The optimality status of a singular trajectory is related to the two semicovariants λ_1 and λ_3. *This status is encoded in X_Y, since $D_3=0$, $D_1D_3>0$ are invariant sets for the solutions of $\dot{v}=X_Y(v)$.*

The proof of Proposition 5.2.9 of [6] is interesting because we compute a normal form in C^0-neighborhood of a singular trajectory γ for the action of G_f on pairs (X, Y).

6. Quadratic control systems. *Classical invariant theory* studies the actions of $\mathrm{GL}(n,\mathfrak{R})$ on the spaces of tensors: vectors, forms, etc. Computing the invariants for such actions is the main problem in analyzing the associated geometries. Thus, numerous algorithms have been described to achieve this task (see, for instance, [16]). We found it interesting to connect this theory with the feedback classification problem by using a specific class of polynomial systems.

DEFINITION 6.1. Consider the set $\mathcal{Q}$ of control systems of $\mathfrak{R}^n$, of the form

$$\dot{x}(t)=Q(x(t))+Bu(t), \tag{12}$$

where $Q=(Q_1,\cdots,Q_n)$, each Q_i being a quadratic form, and B is a constant $n\times m$ matrix whose columns are denoted by $b_1,\cdots,b_m$. They are called *quadratic control systems.*

Consider the subgroup $G_f'\subset G_f$ of triplets(P,α,β), where $P\in\mathrm{GL}(n,\mathfrak{R})$, $\alpha=(\alpha_1,\cdots,\alpha_m)$, each α_i being a quadratic form and β is constant. G_f' is a (Lie group) and acts on a system (Q, B) by the action induced by G_f. The family $\mathcal{Q}$ is *stable* for this action. Let G_d' be the subgroup of elements of the form $(P, 0, \mathrm{id})$ identified with $\mathrm{GL}(n, R)$, and let $G_d'^B$ be the subgroup of $\mathrm{GL}(n,\mathfrak{R})$, leaving invariant the flat distribution generated by $\{b_1,\cdots,b_m\}$, identified to span $\{b_1,\cdots,b_m\}$.

By using an algorithm similar to the one developed in § 3.1, we can compute (generically) the singular extremals.

6.2. Notation. (They are not compatible with § 3.2, but no confusion is possible.) Let Σ be the surface $\{(x,p)\in\mathfrak{R}^{2n}$ subject to $\langle p, B\rangle=0\}$, and let M be the set $\{(x,p)\in\mathfrak{R}^{2n}$ subject to $\langle p, B\rangle=\langle p,[Q,B](x)\rangle=0\}$. Denote by $L(x,p)$ the $m\times 1$ matrix (L_i), where $L_i=\langle p,[Q,[Q,b_i]](x)\rangle$ and let $O(p)$ be the $m\times m$ symmetric matrix (O_{ij}) where $O_{ij}=\langle p,[b_j,[Q,b_i]]\rangle$. Let S be the set $M\cap\mathfrak{R}^n\times\{p\in\mathfrak{R}^n, \det O(p)=0\}$. We denote by $\mathcal{Q}_g$ the set of pairs (Q, B) such that S is a proper subset of M. Take $(Q,B)\in\mathcal{Q}_g$ and let $\hat{u}(x,p)$ be the solution, almost everywhere defined on $\mathfrak{R}^{2n}$, of $L(x,p)+O(x,p)u=0$.

Let $\hat{H}$ be the function almost everywhere defined on $\mathfrak{R}^{2n}$ by $\langle p, Q(x)+B\hat{u}(x,p)\rangle$, and let $\tilde{Z}_{\hat{H}}$ the associated Hamiltonian vector field. The restriction of $\tilde{Z}_{\hat{H}}$ to $M\backslash S$ is denoted by $Z_{\hat{H}}$. A singular extremal of $M\backslash S$ is called of *order* $2m$ (since M is defined by $2m$ equations).

PROPOSITION 6.3. *Let* $(Q, B) \in \mathcal{Q}_g$; *then the singular extremals of order* $2m$ *are the solutions of the* homogeneous *equation*

$$\dot{x} = \frac{\partial \hat{H}}{\partial p}, \qquad \dot{p} = -\frac{\partial \hat{H}}{\partial x}, \qquad (x, p) \in \Sigma. \tag{13}$$

Proof. Compute as in § 3. Observe that the constraints Σ are homogeneous, and since L is quadratic in x and O is constant in x, $\hat{u}(x, p)$ is quadratic in x. Hence $Z_{\hat{H}}$ is homogeneous.

DEFINITION 6.4. Let $\mathcal{Q}_r$ be the set of pairs of $\mathcal{Q}_g$ such that the rank of B is m. Take $(Q, B) \in \mathcal{Q}_r$, then from § 6.2 and Proposition 6.3, the singular extremals in $M\backslash S$ are the solutions of an analytic vector field $Z_{\hat{H}}$. Let λ be the map that associates to $(Q, B) \in \mathcal{Q}_g$ (13). In other words, λ maps (Q, B) onto $(\Sigma, Z_{\hat{H}})$.

The group G'_f acts on pairs $(\Sigma, Z_{\hat{H}})$ by the action induces by G_f. Now observe that the number of equations defining M is $2m$. Hence, by arguments similar to those used in § 4, we have the following.

PROPOSITION 6.5. *The map* λ *is a covariant for the actions of* G'_f *on* $\mathcal{Q}_g$ *and* $\lambda(\mathcal{Q}_g)$.

PROPOSITION 6.6. *Let us assume that* $2m \leq n-1$. *Then for an open dense subset of* $\mathcal{Q}_r$ ($\mathcal{Q}$ *being identified with* $\mathfrak{R}^{((n^2(n+1))/2)+mn}$), *the* G'_f*-classification of pairs* (Q, B) *is equivalent to the* G'_f*-classification of pairs* $(\Sigma, Z_{\hat{H}})$.

Remark 6.7. The only invariant for the G_f-classification of the surfaces Σ is the rank of B, which is equal to m if $(Q, B) \in \mathcal{Q}_g$ and this classification problem is trivial.

7. Quadratic control systems in $\mathfrak{R}^3$.

7.1. Preliminaries. Consider a single input quadratic control system of $\mathfrak{R}^3$:

$$\dot{v}(t) = Q(v(t)) + u(t)b, \tag{14}$$

where $v = (x, y, z)$. As in § 5.2, we set $D_1 = \det(b, [Q, b], [b, [Q, b]])$, $D_2 = \det(b, [Q, b], [Q, [Q, b]])$, and $D_3 = \det(b, [Q, b], Q)$, and we adopt Convention 5.2.1. We consider only pairs (Q, b) such that D_1 is not identically zero on $\mathfrak{R}^3$, and by singular trajectory we mean a singular trajectory contained in $\mathfrak{R}^3\backslash\{D_1 = 0\}$.

From Lemma 5.2.2, the singular trajectories are the solutions of

$$\dot{v} = Q(v) - \frac{D_2(v)}{D_1(v)} b \tag{15}$$

in $\mathfrak{R}^3\backslash\{D_1 = 0\}$.

Let Z be a vector field of $\mathfrak{R}^3$ and let f be a map from $\mathfrak{R}^3$ into $\mathfrak{R}$. The action of G'_f on Z and f is defined by

$$\text{If } (P, \alpha, \beta) \in G'_f, \text{ then } (P, \alpha, \beta) \cdot Z = P * Z \text{ and } (P, \alpha, \beta) \cdot f = f \circ P.$$

By Proposition 5.2.4 and § 6, the G'_f-classification of pairs (Q, b) is equivalent to the G'_f-classification of pairs $(\mathfrak{R}b, Q_b)$, where Q_b designs the vector field on $\mathfrak{R}^3\backslash\{D_1 = 0\}$ defined by (15).

The aim of this section is to investigate the linear classification of vector fields $\{Q_b\}$, to solve the G'_f-classification of pairs (Q, b). *All the computations* (*which are lengthy*) *are omitted since they are detailed in* [2] *and* [5], *and we only present the geometric interpretations of the various semicovariants.*

7.2. Notation. The map $v \to -[Q, b](v)$ is linear and the associated matrix is denoted by A. Let ad A be the adjoint matrix corresponding to A (if A^{-1} exists, then ad $A = \det A \cdot A^{-1}$). Let $w = \operatorname{ad} A(b)$, and observe that $[Q, b](w)$ is colinear to b. Set $L_1 = \mathfrak{R} b$ and $L_2 = \mathfrak{R} w$, and let $\tilde{Q}_b$ be the analytic vector field of $\mathfrak{R}^3$ given by $\tilde{Q}_b = D_1 Q - D_2 b$.

From Lemma 5.2.5, we have the following results.

LEMMA 7.3. *For the G_f'-actions, we have the following* semicovariants:

(i) $\lambda_\pi : (Q, b) \to \tilde{Q}_b$,

(ii) $\lambda_1 : (Q, b) \to D_1$,

(iii) $\lambda_2 : (Q, b) \to D_2$ *restricted to* $D_1 = 0$,

(iv) $\lambda_3 : (Q, b) \to D_3$.

Remark 7.4. The maps λ_1, λ_2, and λ_3 are semicovariants, not covariants. Thus, only the sets $D_1 = 0$, $D_3 = 0$, and $D_1 = 0 \cap D_2 = 0$ have an invariant meaning. This is due to the following property. Since Q_b is homogeneous, with degree 2, then the map $v \to \varepsilon v$, $\varepsilon \in \mathfrak{R} \backslash \{0\}$, transforms Q_b into εQb. Thus, in our classification, we must identify Q_b with εQ_b, and we are dealing, in fact, with *projective geometry*, in which we have no nonconstant polynomial invariants. In particular, the λ_i's cannot be covariants.

7.5. Geometric interpretation of $\tilde{Q}_b$. $\tilde{Q}_b$ is a time reparametrization of the vector field Q_b, similar to the one performed in Remark 4.3. Thus, the solutions of $\dot{v} = \tilde{Q}_b(v)$ in $\mathfrak{R}^3 \backslash \{D_1 = 0\}$, are, up to a reparametrization, singular trajectories. Observe that $\tilde{Q}_b$ is a homogeneous cubic vector field in $\mathfrak{R}^3$.

7.6. Geometric interpretation of $D_1 = 0$. $D_1 = 0$ is the plane generated by $L_1 = \mathfrak{R} b$ and $L_2 = \mathfrak{R} w$ (computations). In particular, $\dot{v} = D_1 Q - D_2 b = -D_2 b$, in $D_1 = 0$. Since b is tangent to $D_1 = 0$, the plane $D_1 = 0$ is then an invariant set for the solutions of $\dot{v} = \tilde{Q}_b(v)$. This set is the union of two types of singularities. First, the set of points where b and $[Q, b]$ are collinear, which are the projections on $\mathfrak{R}^3$ of the singularities of the surface M, defined by $\langle p, b \rangle = \langle p, [Q, b] \rangle = 0$. Secondly, they are the projections onto $\mathfrak{R}^3$ of the set S given by $\langle p, [b, [Q, b]] \rangle = 0 \cap M$.

7.7. Geometric interpretation of $D_3 = 0$. From Definition 5.2.8, $D_3 = 0$ is formed with the exceptional trajectories and according to Proposition 5.2.9, is connected to the time optimality problem. In particular, $D_3 = 0$ is an invariant set for the solution of $\dot{v} = \tilde{Q}_b(v)$. Now observe that D_3 is a cubic form. Linear classification of cubic forms (on $\mathfrak{C}$) is well known. In particular, they have a nontrivial rational invariant, called the modulus. This invariant is very important in our classification problem (see Example 7.10).

7.8. Geometric interpretation of $D_1 = 0 \cap D_2 = 0$. First, observe that the map $v \to D_2(v)$ restricted to $D_1 = 0$ is a cubic form in two variables. By computing, we get that the solutions of $D_1 = 0 \cap D_2 = 0$ are the line $L_2 = \mathfrak{R} w$ and, if a discriminant δ is $\geqslant 0$ (δ is given in [5] and is a semi-invariant), two lines denoted by L_3 and L_4. We have the following nice properties.

Property 7.8.1. In control theory, the system (Q, b) is said to be *weakly controllable* if for all $v \in \mathfrak{R}^3$, $\{Q, b\}_{\text{A.L.}}(v)$ is of rank 3, where $\{Q, b\}_{\text{A.L.}}$ designs the Lie algebra generated by the two vector fields $\{Q, b\}$. This property is clearly feedback invariant and we have that $\{Q, b\}$ is weakly controllable if and only if D_2 is identically zero on $D_1 = 0$, i.e., *the semicovariant λ_2 maps the set on nonweakly controllable pair (Q, b) onto zero.*

Property 7.8.2. Here, we restricted our study to systems (Q, b) such that: span $\{b, [[Q, b], b], [[Q, b], [[Q, b], b]]\} = \mathfrak{R}^3$. By convention, a singular trajectory is a solution of (15), which is defined for $v \in \mathfrak{R}^3 \backslash \{(D_1 = 0\}$. We may ask the following question: when does a system (Q, b) admit singular trajectories $\neq\{0\}$ that are not contained in $\mathfrak{R}^3\backslash\{D_1=0\}$? An associated extremal has to satisfy the relation

$$\langle p, b\rangle = \langle p, [Q, b]\rangle = \langle p, [b, [Q, b]]\rangle = \langle p, [Q, [Q, b]]\rangle = 0,$$

and is then contained in $D_1=0 \cap D_2=0$. Further computations show the following. *There exists a nontrivial trajectory in* $D_1=0$ *if and only if* $L_3=L_4$ (L_3 and L_4 are the two lines previously defined), i.e., $\delta=0$. *Moreover, such a trajectory is supported by the line* $L_3=L_4$. This was predictable, since $D_1=0\cap D_2=0$ is then $L_2 \cup L_3$.

7.9. More semicovariants. Since the equation $\dot v = \tilde Q_b(v)$ is homogeneous, the distribution $v \to \mathfrak{R}\tilde Q_b(v)$ is invariant with respect to the transformations $v \to \varepsilon v$, $\varepsilon \neq 0$ and can be projected onto the sphere S^2. More precisely, if we set $r=(x^2+y^2+z^2)^{1/2}$, $\bar v = v/r$, and $r^2\,dt = d\sigma$, then the equation $dv/dt = \tilde Q_b(v)$ is equivalent to

$$\frac{dr}{d\sigma} = \langle \bar v, \tilde Q_b(\bar v)\rangle r, \tag{16}$$

$$\frac{d\bar v}{d\sigma} = \tilde Q_b(\bar v) - \langle \bar v, \tilde Q_b(\bar v) > \bar v. \tag{17}$$

Equation (17) is defined on S^2 and is called the *projected equation associated with* $\dot v = \tilde Q_b(v)$. *This equation encodes most of behaviors of singular trajectories and has to be carefully studied.*

An outline of this analysis is now given. Let $v_0 \in \mathfrak{R}^3\backslash\{0\}$. One line $\mathfrak{R}v_0$ such that $\tilde Q_b(v_0) = \varepsilon v_0$, $\varepsilon \in \mathfrak{R}$ is called a *ray*. If $\varepsilon = 0$, the ray is a set of singular points for $\dot v = \tilde Q_b(v)$ and if $\varepsilon \neq 0$, the ray is an asymptotic direction for solutions of this equation. Clearly, a ray corresponds bijectively to a singular point for the projected equation.

Now observe that $\tilde Q_b(v)$ can be written as $\tilde Q_b^c(v) + 1/5 \operatorname{div}(\tilde Q_b(v)) \cdot v$, where $\tilde Q_b^c$ is a cubic vector field preserving the standard volume form of $\mathfrak{R}^3$ (i.e., $\operatorname{div} \tilde Q_b^c = 0$). Both $\tilde Q_b$ and $\tilde Q_b^c$ have the same projected equation on S^2 and this decomposition of $\tilde Q_b$ corresponds to the splitting of $\dot v = \tilde Q_b(v)$ into (16) and (17). Moreover, we have the following lemma.

LEMMA. *The following maps are semicovariants*: $\lambda_4 : (Q, b) \to \operatorname{div} \tilde Q_b$, *and* $\lambda_5 : (Q, b) \to \tilde Q_b^c$.

Using the projected equation, the behavior of singular trajectories near $D_1 = 0$ has been classified in [5], in the generic and codimension-one cases. The generic classification is the following. The lines L_i correspond to the only singular points contained in $D_1 = 0$, for the projected equation. They are: *nodes* for $L_1 = \mathfrak{R}b$ and $L_2 = \mathfrak{R}w$, and *saddles* for L_3 and L_4.

Moreover, L_2, L_3, and L_4 are sets of singular points for $\dot v = \tilde Q_b(v)$, and L_1 is an asymptotic direction. Thus, we have the following important result.

PROPOSITION. *For generic pairs* (Q, b) *the distribution* $\mathfrak{R}b$ *is encoded in the vector field* Q_b *(as a ray of* $\tilde Q_b$*). In this case, the* G_f'*-classification of pairs* (Q, b) *is equivalent to the* G_f'*-classification of* Q_b, only.

Example 7.10. Consider the following system:

$$\dot x = a_1 yz + ub_1, \qquad \dot y = a_2 xz + ub_2, \qquad \dot z = a_3 xy + ub_3,$$

where a_1, $a_3 > 0$, $a_2 < 0$, $b = (b_1, b_2, b_3) \in \mathfrak{R}^3$. These equations describe the evolution of the angular velocity in the attitude control problem, when the rigid body is controlled by one torque, whose orientation is given by b, the parameters a_i describing the shape of the satellite. The vector b and in some extent the a_i's can be chosen by the system designer, and we may want to study the feedback classification problem, for this class of systems *when b and the parameters a_i vary.* We have the following results (see [5] for the details).

By computing, we get

$$\begin{aligned}
D_1(v) &= 2a_1b_2b_3(a_3b_2^2 - a_2b_3^2)x + 2a_2b_1b_3(a_1b_3^2 - a_3b_1^2)y \\
&\quad + 2a_3b_1b_2(a_2b_1^2 - a_1b_2^2)z, \\
D_2(v) &= a_3^2b_1b_2(b_1y + b_2x)(a_2x^2 - a_1y^2) \\
&\quad + a_1^2b_2b_3(b_2z + b_3y)(a_3y^2 - a_2z^2) \\
&\quad + a_2^2b_1b_3(b_1z + b_3x)(a_1z^2 - z_3x^2), \\
D_3(v) &= a_2a_3b_1x^2(b_2z - b_3y) + a_1a_3b_2y^2(b_3x - b_1z) \\
&\quad + a_1a_2b_3z^2(b_1y - b_2x).
\end{aligned}$$

The system is weakly controllable if and only if (i) $b_ib_j \neq 0$ for at least one pair (i, j), $i \neq j$, or (ii) $a_3b_1^2 - a_1b_3^2 \neq 0$. Conditions (i) and (ii) have the following interpretation . If (i) is not satisfied, then $D_1 = 0$ is equal to $\mathfrak{R}^3$ and (15) is not defined. If (i) is satisfied, but not (ii), then D_1 is a factor of D_2; this is conformed to Property 7.8.1.

From now on, we assume that (Q, b) is weakly controllable. We can suppose the following:

(1) $a_1 = a_3 = 1$, $a_2 = -1$ (change of coordinates),

(2) $b \in S^2$, i.e., $b_1^2 + b_2^2 + b_3^2 = 1$,

(3) b_1, b_2, $b_3 > 0$ and $b_1 - b_3 > 0$ (symmetries).

With these normalities, we have that

δ is given by $b_1^2b_3^2 - b_2^2b_3^2 - b_1^2b_2^2$ and the lines L_3 and L_4 exist and are distinct if and only if $\delta > 0$, and $L_3 = L_4$ if and only if $\delta = 0$.

The cubic form D_3 is reducible if and only if $b_2 = 0$ or $b_3 = 0$. More precisely, we have that

if $b_3 = 0$, then $D_3 = 0$ is the union of the plane $z = 0$ and the line $x = y = 0$,
if $b_2 = 0$, then D_3 is the union of $y = 0$ and the two planes $x = \pm z$.

We represent below four phase portraits for the projected equation, defined on S^2. They are obtained by direct computations ((i) and (iv)) and bifurcation analysis ((ii) and (iii)). The plane $D_1 = 0$, generated by L_1 and L_2, is identified with the equator, and we draw only the phase portrait in the northern hemisphere (the phase portrait in the southern hemisphere can be deduced by symmetry).

Cases (i) *and* (ii) (Fig. 1). Comments. Case (i) represents the situations $b = (b_1, 0, b_3)$, and case (ii) $b = (b_1, \varepsilon, b_3)$, $\varepsilon \neq 0$ and small. For both cases, L_3 and L_4 exist and are distinct. A separatrix cycle $O'L_3L_4$ in (i) is transformed in (ii) into a limit cycle corresponding to one loop of the cubic $D_3 = 0$ projected on S^2.

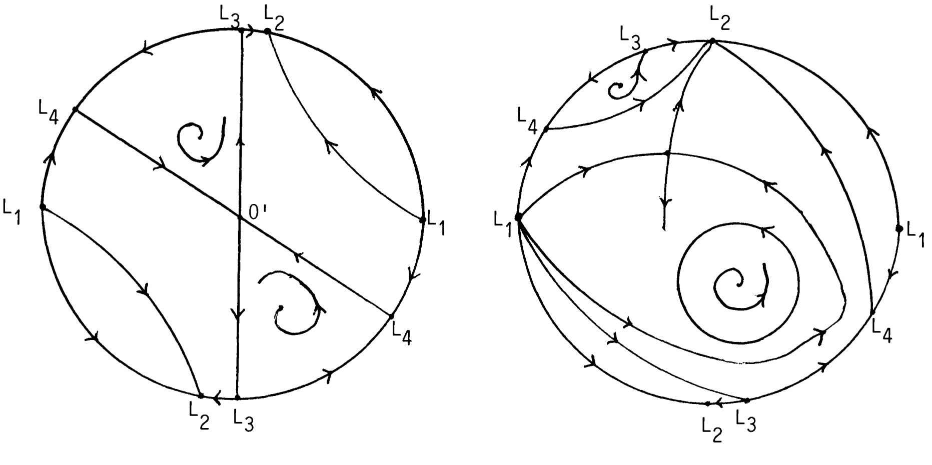
Cases (iii) *and* (iv) (Fig. 2). Comments. Case (iv) represents the situations $b = (b_1, b_2, 0)$, and (iii) $b = (b_1, b_2, \varepsilon)$, $\varepsilon \neq 0$ and small. In both cases, the two lines L_3 and L_4 do not exist. The transition (iii) → (iv) corresponds to the vanishing of a small limit cycle by Hopf bifurcation. This cycle is still a loop of $D_3 = 0$ projected on S^2.

The intermediary cases between (ii) and (iii) are described in [5].

Conclusion. Systems (Q, b) corresponding to cases (i)-(iv) are not feedback equivalent: *the phase portraits of $\tilde{Q}_b$ are not even C^0-equivalent.*

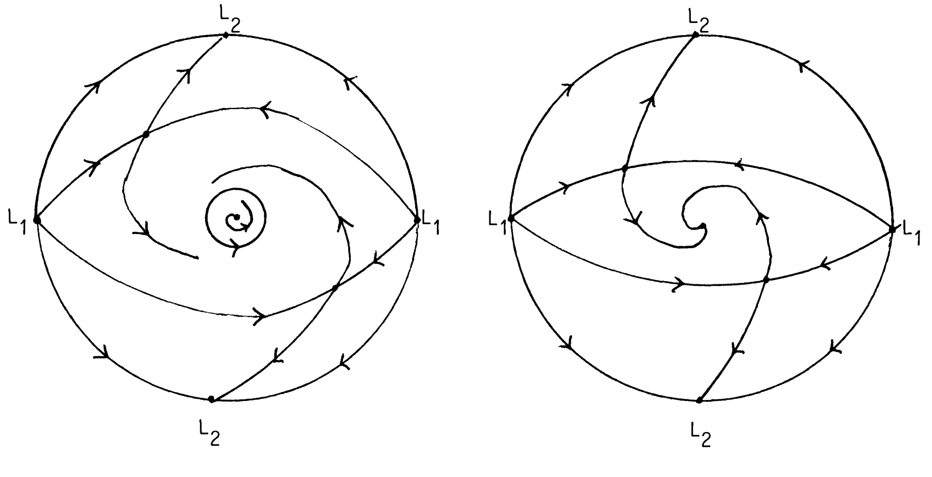
Now consider a quadratic system of $\mathfrak{R}^3$, with two inputs:

$$\dot{v}(t) = Q(v(t)) + Bu(t). \tag{18}$$



(i) (ii)

FIG. 1



(iii) (iv)

FIG. 2

Assume that rank $B = 2$, and let b_1 and b_2 be the columns of B. Let $L = \bigcap_{i=1,2}\{v \in \Re^3;\ \det(b_1, b_2, [Q, b_i](v)) = 0\}$. From § 6, the singular trajectories are contained in the vector space L, and we have the following canonical forms for the G_f'-classification.

PROPOSITION 7.11. *(Q, B) is G_f'-equivalent to*

$$Q_1(v)\frac{\partial}{\partial x} + u_1\frac{\partial}{\partial y} + u_2\frac{\partial}{\partial z},$$

where Q_1 has one of the following forms in Table 1.

The proof follows from straightforward computations. The remarkable property is the following. For orbits (a) and (b), the line L supporting the singular trajectories is precisely the one to choose to complete the linear span of $\{b_1, b_2\}$ to get the basis

TABLE 1

Normal forms	Geometric interpretation
(a) Generic orbits	dim $L=1$, $\Re^3=L+\text{Span } B$
(a_1) $Q_1=x^2+y^2+z^2$	Singular flow (on L): $\rightarrow\!-\!\overset{0}{\bullet}\!\rightarrow\!\rightarrow$
(a_2) $Q_1=x^2-y^2-z^2$	Optimality status of the singular trajectoires ($\neq 0$)
(a_3) $Q_3=x^2+y^2-z^2$	(a_1) = fast, (a_2) = slow, (a_3) = locally controllable
(b_1) $Q_1=y^2+z^2$	dim $L=1$, $\Re^3=L+\text{Span } B$
(b_2) $Q_1=y^2-z^2$	Singular flow: $\bullet\!-\!\bullet\!-\!\bullet\!-\!\bullet\!-\!\overset{0}{\bullet}\!-\!\bullet\!-\!\bullet\!-\!\bullet\!-\!\bullet\!\rightarrow$
	for (b_2) each singular trajectory is locally controllable, contrarily to (b_1)
(c) $Q_1=xz+y^2$	dim $L=1$, $L\subset \text{Span } B$
(d_1) $Q_1=x^2+y^2$	dim $L=2$, dim $(L\cap \text{Span } B)=1$
(d_2) $Q_1=x^2-y^2$	$\Re^3=L+\text{span } B$
	Singular flow: $\dot{x}=\varepsilon x^2$, $\dot{z}=u_2$
(d_3) $Q_1=y^2$	Orbits separated by nature of the singular flow and optimality status
(e) (Q, b) not weakly controllable	
(e_1) $Q_1=0$	$L=\Re^3$. Singular flow: $\dot{x}=0$, $\dot{y}=u_1$, $\dot{z}=u_2$
(e_2) $Q_1=x^2$	$L=\Re^3$. Singular flow: $\dot{x}=x^2$, $\dot{y}=u_1$, $\dot{z}=u_2$
(e_3) $Q_1=xy$	$L=\text{span } B$

for our canonical form. This classification for generic orbits, i.e., with maximal dimensions, is generalized to $\Re^n$, when $m=n-1$, in [3].

7.12. G_f-feedback equivalence with quadratic control systems. For simplicity, consider a single input system *in* $\Re^3$ denoted by (X, Y). We may ask the question, when is this system G_f-equivalent to a *quadratic* control system (Q, b)? (This problem can also be locally studied.) For such an equivalence, the distribution $\Re Y$ has to be equivalent to a flat distribution. Moreover, from Proposition 5.2.4, the vector field denoted by X_Y, whose solutions are the singular trajectories of (X, Y) contained in $\Re^3 \backslash \{D_1^{X,Y}=0\}$, must be, up to a change of time parameter, diffeomorphic to the homogeneous cubic vector field $\tilde{Q}_b$ introduced in § 7.2.

Thus, our problem is tightly connected with the C^ω-equivalence of a differential equation with a polynomial one. The local version of this problem is a standard question, extensively studied. Near a regular point, any vector field is diffeomorphic to $\partial/\partial x$. Near a singular point, this problem is very difficult. Relevant contributions were made by Lyapunov, Poincaré, and Dulac (see [1]), and this problem is still the object of contemporary studies (see [24] and [26], etc).

The G_f-equivalence of (X, Y) with a quadratic control system (Q, b) can also be solved using the G_f'-canonical forms computed in [3].

8. Conclusion. In this conclusion, we briefly indicate how to complete the results obtained in this article. First, note that our theory is global, in the sense that the singular flow is, up to a singular set, globally defined. On the other hand, with the trick introduced in Remark 4.3, we can delete this singular set. This is useful in studying the C^∞ case.

There is a gap in our aim to identify the feedback classification with the classification of pairs $(\Sigma, Z_{\hat{H}})$. We must assume that the singular trajectories are not contained in a proper subvariety of $\Re^n$. However, the analysis done in §§ 5.1 and 7.1 indicates how to avoid this assumption. The *position* of this subvariety with respect to the distribution Δ_Y and the *optimality status* of the singular trajectories allow us to carry out the classification. This position is encoded in pairs $(\Sigma, Z_{\hat{H}})$, but not the optimality status. The time optimality status is an invariant of the time minimal control problem. Therefore, we conjecture that for generic systems (X, Y) when the feedback classification is not trivial, i.e., when $m < n$, then the feedback classification is equivalent to the classification of the time minimal control problem. Under generic assumptions, the optimality status can be deduced from the so-called Legendre–Clebsch condition arising from the high-order maximal principle [21]. This can be applied to analyze the case $m = n - 1$, n odd.

Now a nontrivial extension of this article would be the analysis of the feedback equivalence problem, for general systems given by $\dot{x}(t) = f(x(t), u(t))$, where $x \in \Re^n$, $u \in \Re^m$, and f is smooth. The group action being extended to transformations of the form $x = \psi(y)$, $u' = \Phi(x, u)$, where the maps $y \to \psi(y)$ and $u \to \Phi(\cdot, u)$ are diffeomorphisms. Now observe that even in this case, the singularities of the input-state mapping are well defined. From § 2, it is clear that the maximum principle is still the key tool to investigate the feedback equivalence problem. Let us illustrate this assertion by an example.

Recall that the Hamiltonian $\tilde{H}$ is defined by $\tilde{H}(x, p, u) = \langle p, f(x, u)\rangle$.

Now observe that the surface Σ defined by $\partial\tilde{H}/\partial u = 0$ still has an invariant meaning for our classification problem. Consider the restriction of the Hessian matrix $(\partial^2\tilde{H})/(\partial u_i \partial u_j)$ to the surface Σ. This matrix is the intrinsic derivative and its rank is an invariant of the feedback classification problem. It has the following interpretation.

Let us assume the rank maximal, i.e., equal to m. Then the singular control can be (locally) computed as a map $\hat{u} : (x, p) \to \Re^m$, by using the implicit function theorem to solve the equation $\partial\tilde{H}/\partial u = 0$. Conversely, if the restriction of the Hessian matrix to Σ is identically zero, a straightforward analysis shows that, with regularity assumptions on the map $u \to f(x, u)$, then the system is feedback equivalent to an affine system. The vanishing of the Hessian matrix Σ being equivalent to say that the image of $\Re^m$ by the map $u \to f(x, u)$ is a flat submanifold of $\Re^m$. In other words, the only difference between the affine and nonaffine case is when solving the equation $\partial\tilde{H}/\partial u = 0$.

To summarize, the key to providing a general theory of the feedback classification is *Pontryagin's Maximum Principle.*

REFERENCES

[1] V. I. ARNOLD, *Chapitres supplémentaires de la théorie des équations différentielles*, Ed. Mir, Moscow, 1980.

[2] B. BONNARD, *On singular extremals in the time minimal control problem in* $\Re^3$, SIAM J. Control Optim., 23 (1985), pp. 794–802.

[3] ———, *Quadratic control systems*, Math. Control Signals Systems, 4 (1991), pp. 139–160.

[4] ———, *Generic properties for singular trajectories*, in Proc. 27th CDC/IEEE Conference, Austin, TX, 1988.

[5] ———, *Contribution à l'étude des trajectoires singulières*, Report L.A. Grenoble, 1987.

[6] B. BONNARD AND I. KUPKA, *Feedback equivalence and the time optimality problem*, Forum Math., 1991, to appear.

[7] R. W. BROCKETT, *Feedback invariants for nonlinear systems*, in Proc. IFAC Congress, Helsinki, 1978.

[8] P. BRUNOVSKY, *A classification of linear controllable systems*, Kybernetika, 6 (1970), pp. 173–183.

[9] E. CARTAN, *Oeuvres complètes*, Vol. 2, Gauthier-Villars, Paris, 1953.

[10] ———, *Leçons sur la géométrie des espaces de Riemann*, Gauthier-Villars, Paris, 1963.

[11] J. A. DIEUDONNÉ AND J. B. CARREL, *Invariant Theory: Old and New*, Academic Press, New York, 1971.

[12] R. B. GARDNER, *Differential geometric methods interfacing control theory*, in Differential Geometric Control Theory, Brockett, Millman, and Sussmann, eds., Birkhäuser, Boston, 1983.

[13] R. B. GARDNER, W. F. SHADWICK, AND G. R. WILKENS, *A geometric isomorphism with application to closed loop controls*, preprint, 1988.

[14] E. G. GILBERT, *Functional expansion for the response of nonlinear differential systems*, IEEE Trans. Automat. Control, 22 (1977), pp. 909-921.

[15] C. GODBILLON, *Géometrie différentielle et mécanique analytique*, Herman collection Méthodes, Paris, 1969.

[16] G. B. GUREVITCH, *Foundations of the theory of algebraic invariants*, Noordhoff, Groningen, the Netherlands, 1964.

[17] H. HERMÈS AND J. P. LASALLE, *Functional Analysis and Time Optimal Control*, Mathematics in Sciences and Engineering, Vol. 56, Academic Press, New York, 1969.

[18] L. R. HUNT, R. SU, AND G. MEYER, *Global transformation of nonlinear systems*, IEEE Trans. Automat. Control, 28 (1983), pp. 24-31.

[19] B. JAKUBCZYCK AND W. RESPONDEK, *On linearization of control systems*, Bull. Acad. Pol. Sci., 28 (1980), pp. 517-522.

[20] A. KOLMOGOROV AND S. FOMINE, *Eléments de la théorie des fonctions et de l'analyse fonctionnelle*, Ed. Mir, Moscow, 1974.

[21] A. J. KRENER, *The high order maximal principle and its applications to singular extremals*, SIAM J. Control Optim., 15 (1977), pp. 256-293.

[22] E. B. LEE AND L. MARKUS, *Foundations of optimal control theory*, SIAM Series in Applied Math., John Wiley, New York, 1967.

[23] S. LEFSCHETZ, *Differential Equations: Geometric Theory*, Dover, New York, 1977.

[24] J. MARTINET AND J. P. RAMIS, *Problèmes de modules pour des équations différentielles non linéaires du premier ordre*, Publication Mathématiques Institut des Hautes Etudes Scientifiques, 55 (1982), pp. 63-164.

[25] S. STERNBERG, *Differential geometric*, Chelsea, New York, 1983.

[26] F. TAKENS, *Singularities of vector fields*, Publication Mathématiques Institut des Hautes Etudes Scientifiques, 43 (1973), pp. 47-100.

SICON 503
SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 6, pp. 1322-1347, November 1991

SOME CHARACTERIZATIONS OF OPTIMAL TRAJECTORIES IN CONTROL THEORY*

PIERMARCO CANNARSA† AND HALINA FRANKOWSKA‡

Abstract. Several characterizations of optimal trajectories for the classical Mayer problem in optimal control are provided. For this purpose the regularity of directional derivatives of the value function is studied: for instance, it is shown that for smooth control systems the value function V is continuously differentiable along an optimal trajectory $x:[t_0, 1]\to \mathbf{R}^n$ provided V is differentiable at the initial point $(t_0, x(t_0))$.

Then the upper semicontinuity of the optimal feedback map is deduced. The problem of optimal design is addressed, obtaining sufficient conditions for optimality. Finally, it is shown that the optimal control problem may be reduced to a viability one.

Key words. Hamilton-Jacobi equation, optimal synthesis, semiconcave function, viability theory, viscosity solutions, set-valued derivatives, sufficient conditions for optimality

AMS(MOS) subject classifications. 49K15, 49L05, 93B50, 93C15

1. Introduction.

Consider the optimal control problem

$$\text{minimize } g(x(1))$$

over all solutions of the control system

$$x'=f(t, x, u(t)), \qquad u(t)\in U \tag{1}$$

satisfying the initial condition

$$x(0)=\xi_0. \tag{2}$$

We recall that by a simple change of variables the classical Bolza problem in control theory

$$\text{minimize}\left\{\varphi(x(1))+\int_0^1 L(t, x(t), u(t))\,dt\right\}$$

over the trajectory control pairs (x, u) of (1), (2) may be reduced to the one under consideration.

The goal of the optimal control theory is to find necessary and sufficient conditions for optimality and to construct optimal trajectories. Several results establishing necessary conditions are available in the form of the maximum principle. In this paper, we show that additional information (including sufficient conditions for optimality, optimal design, and optimal synthesis) may be obtained from some properties of the value function, which is defined by

$$V(t_0, x_0)=\inf\{g(x(1))\,|\,x \text{ is a solution of (1) on } [t_0, 1], x(t_0)=x_0\}.$$

In general, even in very regular situations, the value function is not differentiable. Nevertheless, we prove in this paper that the differentiability of V is preserved along

* Received by the editors September 8, 1989; accepted for publication (in revised form) October 9, 1990.

† Dipartimento di Matematica Università di Pisa, via F. Buonarroti 2, 56127 Pisa, Italy. The work of this author was partially supported by the national project "Equazioni di Evoluzione e Applicazioni Fisico-Matematiche" (MPI, 40%).

‡ CEREMADE (Centre de la Recherche de Mathématique de la Décision), Université Paris-Dauphine, 75775 Paris, Cx 16, France, and IIASA, International Institute for Applied Systems Analysis, Laxenburg, A2361 Austria.

optimal trajectories. More precisely, we show that if V is differentiable at some point (t_0, x_0) and $\bar{x}$ denotes any optimal solution starting from x_0 at time t_0, then V is differentiable at $(t, \bar{x}(t))$ for every $t \in [t_0, 1]$ (see Corollary 5.3). Actually, the derivative $-V'_x(t, \bar{x}(t))$ is equal to the co-state of the Pontriagin maximum principle, which we recall in § 4. In the same section, we also derive some inclusions connecting the co-state with the superdifferential of the value function. Among these results, Theorem 4.2 is related to [25] and Proposition 4.4 to [26, Prop. 3.1].

When the Hamiltonian H is smooth enough and the value function is differentiable at $(0, \xi_0)$, then the following necessary and sufficient condition for optimality holds true. Let $x(\cdot), p(\cdot)$ solve the Hamiltonian system

$$\begin{aligned} x'(t) &= \frac{\partial H}{\partial p}(t, x(t), p(t)), \\ p'(t) &= -\frac{\partial H}{\partial x}(t, x(t), p(t)), \qquad t \in [0, 1]. \end{aligned} \tag{3}$$

Then x is optimal if and only if $x(0) = \xi_0$, $p(0) = -V'_x(0, \xi_0)$ (see Theorem 4.6 for more general statements.)

Even when the Hamiltonian is not smooth, the value function may still be used to construct the optimal feedback map

$$G(t, x) = \left\{ v \in f(t, x, U) \,\middle|\, \frac{\partial V}{\partial(1, v)}(t, x) = 0 \right\}. \tag{4}$$

In fact, the following property holds true: a trajectory $\bar{x}$ of (1) is optimal for our optimization problem if and only if it is a solution of the differential inclusion

$$x' \in G(t, x), \qquad x(0) = \xi_0. \tag{5}$$

We refer to [17], [6] for some developments in this direction.

To investigate regularity properties of the set-valued map G, we prove the existence of the directional derivatives of V. For this aim we show that, under very general assumptions on the control system, the value function is semiconcave (see Theorem 5.1).

As a consequence of the semiconcavity of V, we obtain that the feedback map G is upper semicontinuous and has nonempty compact images (see Theorem 6.1).

In particular, whenever the feedback map G is single-valued, it is continuous. From the above it follows that in this special case, optimal trajectories are continuously differentiable.

Moreover, if the data are convex, then G has convex values and the inclusion (5) fits the well-investigated framework of upper semicontinuous convex-valued maps. When the map G does not have convex images, the above characterization of optimal trajectories is not easy to apply. To overcome this difficulty, we provide an alternative approach based on viability theory. More precisely, we observe that solving the optimal control problem is equivalent to solving a control system with state constraints:

(i) $t' = 1,$

(ii) $x' = f(t, x, u), \qquad u \in U,$

(iii) $z' = 0,$

(iv) $(t, x(t), z(t)) \in \text{Graph}(V),$

(v) $t(0) = 0, \quad x(0) = \xi_0, \quad z(0) = V(0, \xi_0).$

The last system is a viability problem and may be approached using many results of viability theory (see [20], [2], [1], and references contained therein). We underline that in this case dynamics (i)–(iii) remain regular, but we have to keep trajectories in the set Graph (V) according to the relation (iv) (see [18]).

Finally, we treat the case involving the endpoint constraints ($x(1)\in K_1$) via penalization techniques. We prove that the value function of such a problem may be approximated by the value function of problems with free endpoints (see Theorem 8.1). A similar result holds true for optimal trajectories.

Some of the results of the present paper were announced in [7]. Moreover, this approach may be extended to infinite-dimensional control problems, as we show in [8].

The plan of the paper is as follows. Section 2 contains basic material on the value function. In § 3 we recall some definitions of set-valued gradients and investigate properties of semiconcave functions. Necessary and sufficient conditions for optimality are described in § 4, while § 5 is devoted to the semiconcavity of the value function. The optimal feedback map is studied in § 6 and viability theory is applied to optimal trajectories in § 7. In § 8 we address the problem with endpoint constraints.

2. Value function in optimal control. Consider a complete separable metric space U and a continuous function

$$f:[0,1]\times\mathbf{R}^n\times U\to\mathbf{R}^n.$$

We associate with it the control system

$$x'(t)=f(t,x(t),u(t)),\qquad u(t)\in U\quad\text{a.e.}\tag{6}$$

An absolutely continuous function $x:[t_0,t_1]\to\mathbf{R}^n$ is called a trajectory of (6) if there exists a measurable function $u:[t_0,t_1]\to U$ such that $x'(t)=f(t,x(t),u(t))$ almost everywhere in $[t_0,t_1]$.

Let $g:\mathbf{R}^n\to\mathbf{R}$ and $\xi_0\in\mathbf{R}^n$ be given. We investigate the minimization problem

$$\text{minimize}\,\{g(x(1))\,|\,x\text{ is a solution of (6) on }[0,1],x(0)=\xi_0\}.\tag{7}$$

The dynamic programming approach associates with this problem the value function defined by

$$V(t_0,x_0)=\inf\{g(x(1))\,|\,x\text{ is a solution of (6) on }[t_0,1],x(t_0)=x_0\}.\tag{8}$$

Throughout the whole paper we impose the following assumptions:

$$\begin{array}{ll}\text{(i)} & f\text{ is continuous in }[0,1]\times\mathbf{R}^n\times U,\\ \text{(ii)} & \exists k\in L^1(0,1;\mathbf{R}_+),\ \forall(t,u)\in[0,1]\times U,\ f(t,\cdot,u)\text{ is }k(t)\text{-Lipschitz},\\ \text{(iii)} & \exists\gamma>0\text{ such that }\forall(t,u)\in[0,1]\times U,\ \|f(t,x,u)\|\leqq\gamma(\|x\|+1),\\ \text{(iv)} & g\text{ is locally Lipschitz.}\end{array}\tag{9}$$

Our assumptions allow us to apply the relaxation theorem from [2] to show that V is actually equal to the value function of the relaxed problem in which (6) is replaced by the differential inclusion

$$x'(t)\in\overline{\text{co}}\,f(t,x(t),U)\quad\text{a.e.}\tag{10}$$

We recall that an absolutely continuous function $x:[t_0,t_1]\mapsto\mathbf{R}^n$ is called a trajectory of (10) if for almost every $t\in[t_0,t_1]$, $x'(t)\in\overline{\text{co}}\,f(t,x(t),U)$. Hence we will study the following minimization problem:

$$\text{minimize}\,\{g(x(1))\,|\,x\text{ is a solution of (10) on }[0,1],x(0)=\xi_0\}.\tag{11}$$

The corresponding value function is given by

$$V^{\text{co}}(t_0,x_0)=\inf\{g(x(1))\,|\,x\text{ is a solution of (10) on }[t_0,1],x(t_0)=x_0\}.$$

THEOREM 2.1. *Assume* (9). *Then, for all* $(t_0,x_0)\in[0,1]\times\mathbf{R}^n$ *we have*

$$V(t_0,x_0)=V^{\text{co}}(t_0,x_0)=\min\{g(x(1))\,|\,x\textit{ is a solution of }(10)\textit{ on }[t_0,1],x(t_0)=x_0\}.$$

Proof. From the relaxation and the parametrization theorems (see [2]), we know that the closure in the metric of uniform convergence of trajectories of (6) defined on the time interval $[t_0, 1]$ and starting at x_0, which is compact in $\mathscr{C}([t_0, 1]; \mathbf{R}^n)$), is equal to the set of trajectories of (10) starting at x_0 and defined on $[t_0, 1]$. This ends the proof. □

It is well known that the value function is nondecreasing along trajectories of (6) and therefore a trajectory $x:[t_0, 1] \to \mathbf{R}^n$ satisfies $V(t_0, x(t_0)) = g(x(1))$ if and only if $V(t, x(t)) \equiv g(x(1))$. This leads to a verification technique in optimal control:

> A trajectory $x:[0, 1] \to \mathbf{R}^n$ of the control system (6) is optimal for the problem (7) if and only if $x(0) = \xi_0$ and $V(t, x(t)) \equiv \text{const}$ (in this case $V(t, x(t)) \equiv g(x(1))$).

Hence, instead of looking for an optimal trajectory for the problem (7), we can search for a trajectory of (6) satisfying the initial condition and such that the value function is constant along it.

We recall that the directional derivative of a function $\varphi: \mathbf{R}^n \to \mathbf{R}$ at $x_0 \in X$ in the direction $\Theta \in X$ (when it exists) is defined by

$$\frac{\partial \varphi}{\partial \Theta}(x_0) = \lim_{h \to 0+} \frac{\varphi(x_0 + h\Theta) - \varphi(x_0)}{h}.$$

PROPOSITION 2.2. *Assume* (9). *Then the value function V is locally Lipschitz. Furthermore, for every trajectory x of* (6) *on* $[0, 1]$ *and for almost every* $t \in [0, 1]$, *there exists the directional derivative*

$$\frac{\partial V}{\partial(1, x'(t))}(t, x(t)).$$

Proof. The local Lipschitz continuity of V is well known. It can be checked by arguments similar to [16, Thm. 4.2, p. 85] (see also [17]).

Fix a trajectory $x(\cdot)$. Then the function $t \to \varphi(t) := V(t, x(t))$ is absolutely continuous. Fix t such that φ and x are differentiable at t. Then

$$\lim_{h \to 0+} \frac{V(t+h, x(t) + hx'(t)) - V(t, x(t))}{h} = \lim_{h \to 0+} \frac{V(t+h, x(t+h)) - V(t, x(t))}{h}$$

and the proof follows. □

When the value function is directionally differentiable, it has many properties related to the dynamics of the system.

PROPOSITION 2.3. *Assume* (9). *If for some* $(t_0, x_0) \in [0, 1[\times \mathbf{R}^n$ *and* $v \in \overline{\text{co}}\, f(t_0, x_0, U)$, *V has the directional derivative at* (t_0, x_0) *in the direction* $(1, v)$, *then this directional derivative is nonnegative.*

Proof. Consider a solution $x(\cdot)$ of the differential inclusion (10) satisfying $x(t_0) = x_0$, $x'(t_0) = v$ (by [2] such solution does exist). Since V is locally Lipschitz at (t_0, x_0) and nondecreasing along trajectories of (10) (thanks to Theorem 2.1), we obtain

$$\lim_{h \to 0+} \frac{V(t_0 + h, x_0 + hv) - V(t_0, x_0)}{h} = \lim_{h \to 0+} \frac{V(t_0 + h, x(t_0 + h)) - V(t_0, x_0)}{h} \geqq 0. \quad \square$$

To characterize optimal trajectories, we introduce two following feedback maps $G: [0, 1] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ and $G^{\text{co}}: [0, 1] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ defined, respectively, by

$$G(t, x) = \left\{ v \in f(t, x, U) \,\middle|\, \frac{\partial V}{\partial(1, v)}(t, x) = 0 \right\}$$

and

$$G^{co}(t,x)=\left\{v\in\overline{\mathrm{co}}\,f(t,x,U)\;\middle|\;\frac{\partial V}{\partial(1,v)}(t,x)=0\right\}$$

(note that the sets $G(t,x)$ and $G^{co}(t,x)$ may be empty).

Then we have the following characterizations of optimal trajectories.

THEOREM 2.4. *Assume* (9). *Then the following two statements are equivalent*:

(i) *x is a trajectory of the differential inclusion*

$$x'\in G(t,x) \tag{12}$$

defined on the time interval $[t_0,1]$.

(ii) *x is a trajectory of the control system* (6) *defined on the time interval* $[t_0, 1]$ *and, for every* $t\in[t_0,1]$, $V(t,x(t))=g(x(1))$.

For the relaxed system (10) *the following two statements are equivalent*:

(iii) *x is a trajectory of the differential inclusion*

$$x'\in G^{co}(t,x) \tag{13}$$

defined on the time interval $[t_0,1]$.

(iv) *x is a trajectory of the differential inclusion* (10) *defined on the time interval* $[t_0,1]$ *and, for every* $t\in[t_0,1]$, $V(t,x(t))=g(x(1))$.

Proof. Fix a trajectory x of (12) defined on time interval $[t_0,1]$ and set $\varphi(t)=V(t,x(t))$ for every $t\in[t_0,1]$. Since V is locally Lipschitz (recall Proposition 2.2), φ is absolutely continuous. Thus, by Proposition 2.2, for almost all $t\in[t_0,1]$,

$$\varphi'(t)=\frac{\partial V}{\partial(1,x'(t))}(t,x(t)).$$

By Filippov's lemma $x(\cdot)$ is a solution to (6) (see [2]). Thus $\varphi'(t)=0$ almost everywhere in $[t_0,1]$. Consequently, φ is constant and is equal to $V(1,x(1))=g(x(1))$. Assume next that (ii) holds true. Then differentiating the map $t\to\varphi(t)$, we obtain that for every $t\in\,]t_0,1[$, $\varphi'(t)=0$. Thus

$$\frac{\partial V}{\partial(1,x'(t))}(t,x(t))=0$$

almost everywhere and therefore for almost all $t\in[t_0,1]$, $x'(t)\in G(t,x(t))$. The proof of the second statement is analogous and is omitted. □

COROLLARY 2.5. *Assume* (9). *Then a trajectory* $x:[0,1]\to\mathbf{R}^n$ *is an optimal solution of problem* (7) *if and only if it is a solution of the differential inclusion* (12) *and* $x(0)=\xi_0$. *An analogous statement holds true for the relaxed problem* (11) *and the differential inclusion* (13).

Proof. Since V is nondecreasing along trajectories of the control system (6), we deduce that $\bar{x}(\cdot)$ is optimal for the control problem (7) if and only if V is constant along $\bar{x}$. Theorem 2.4 ends the proof. □

THEOREM 2.6. *Assume* (9). *Then for every* $t_0\in[0,1]$, $x_0\in\mathbf{R}^n$, *inclusion* (13) *has at least one solution satisfying* $x(t_0)=x_0$.

Proof. Consider the optimal control problem

$$\text{minimize } g(x(1))$$

over the solutions of the differential inclusion

$$x'(t)\in\overline{\mathrm{co}}\,f(t,x(t),U)\quad\text{a.e. in }[t_0,1],\ x(t_0)=x_0.$$

By Theorem 2.1, it has at least one optimal solution $\bar{x}$. Furthermore, $V(t, \bar{x}(t)) \equiv g(\bar{x}(1))$. Theorem 2.4 ends the proof. □

3. Some preliminaries on nonsmooth functions. We denote by B the closed unit ball in $\mathbf{R}^n$ and by $B_r(x_0)$ the closed ball in $\mathbf{R}^n$ with radius r and center at x_0. Consider an open set $\Omega \subset \mathbf{R}^n$ and a function $\varphi : \Omega \to \mathbf{R}$. When φ is not differentiable, it is possible to define its gradient taking weaker limits of differential quotients.

DEFINITION 3.1. Let $x_0 \in \Omega$. The *superdifferential* of φ at x_0 is the closed convex set defined as follows:

$$D^+\varphi(x_0) = \left\{ p \in \mathbf{R}^n \;\middle|\; \limsup_{x \to x_0} \frac{\varphi(x) - \varphi(x_0) - \langle p, x - x_0 \rangle}{\|x - x_0\|} \leqq 0 \right\},$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product.

The *subdifferential* is defined in a similar way:

$$D^-\varphi(x_0) = \left\{ p \in \mathbf{R}^n \;\middle|\; \liminf_{x \to x_0} \frac{\varphi(x) - \varphi(x_0) - \langle p, x - x_0 \rangle}{\|x - x_0\|} \geqq 0 \right\}.$$

It is not difficult to show that φ is Fréchet differentiable at x_0 if and only if both the super- and the subdifferential of φ are not empty at x_0. In this case

$$D^+\varphi(x_0) = D^-\varphi(x_0) = \{\varphi'(x_0)\},$$

where $\varphi'(x_0)$ denotes the gradient of φ at x_0. We always have $D^+\varphi(x_0) = -D^-(-\varphi)(x_0)$.

The super- and subdifferential may also be characterized using the *Dini* directional derivatives, which are defined in the following definition.

DEFINITION 3.2. The *lower Dini derivative* of φ at x_0 in the direction Θ is given by

$$\partial^-\varphi(x_0)(\Theta) = \liminf_{h \to 0+, \Theta' \to \Theta} \frac{\varphi(x_0 + h\Theta') - \varphi(x_0)}{h}$$

and the *upper Dini derivative* of φ at x_0 in the direction Θ is defined by

$$\partial^+\varphi(x_0)(\Theta) = \limsup_{h \to 0+, \Theta' \to \Theta} \frac{\varphi(x_0 + h\Theta') - \varphi(x_0)}{h}. \tag{14}$$

Clearly,

$$\partial^-\varphi(x_0) = -\partial^+(-\varphi)(x_0). \tag{15}$$

When φ is Lipschitz at x_0, then the definition may be simplified as follows:

$$\partial^-\varphi(x_0)(\Theta) = \liminf_{h \to 0+} \frac{\varphi(x_0 + h\Theta) - \varphi(x_0)}{h}$$

and

$$\partial^+\varphi(x_0)(\Theta) = \limsup_{h \to 0+} \frac{\varphi(x_0 + h\Theta) - \varphi(x_0)}{h}.$$

From [17, Lemma 2.7] (see also [3, Chap. 6]) we know that

$$D^-\varphi(x_0) = \{ p \in \mathbf{R}^n \mid \forall \Theta \in \mathbf{R}^n, \partial^-\varphi(x_0)(\Theta) \geqq \langle p, \Theta \rangle \}$$

and

$$D^+\varphi(x_0) = \{ p \in \mathbf{R}^n \mid \forall \Theta \in \mathbf{R}^n, \partial^+\varphi(x_0)(\Theta) \leqq \langle p, \Theta \rangle \}. \tag{16}$$

DEFINITION 3.3. Assume that φ is Lipschitz at $x_0 \in \Omega$. The *regularized lower derivative* of φ at x_0 in the direction $\Theta \in \mathbf{R}^n$ is defined by

$$\varphi^o_-(x_0, \Theta) = \liminf_{h \to 0+, x \to x_0} \frac{\varphi(x+h\Theta) - \varphi(x)}{h}.$$

This notion is a "lower version" of Clarke's definition of directional derivative. Indeed, it can be easily checked that

$$(17) \qquad \varphi^o_-(x_0, \Theta) = -\varphi^o(x_0, -\Theta) = -(-\varphi)^o(x_0, \Theta),$$

where $\varphi^o(x_0, \Theta)$ denotes the directional derivative from [11].

PROPOSITION 3.4. *Let $\varphi : \mathbf{R}^n \to \mathbf{R}$ be Lipschitz at $x_0 \in \mathbf{R}^n$. Then the function $\Theta \to \varphi^o_-(x_0, \Theta)$ is concave.*

This result may be deduced from [11, Prop. 2.1.1].

We investigate next the closedness of the level sets of the regularized lower derivative.

PROPOSITION 3.5. *Let $\varphi : \mathbf{R}^n \to \mathbf{R}$ be a locally Lipschitz function and define the set-valued map $Q : \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ by*

$$Q(x) = \{\Theta \in \mathbf{R}^n \mid \varphi^o_-(x, \Theta) \leqq 0\}.$$

Then Q has nonempty closed images and the graph of the map Q is closed.

Proof. Clearly, for every x, $0 \in Q(x)$. It remains to show that for every sequence $(x_n, \Theta_n) \in \mathbf{R}^n \times \mathbf{R}^n$ converging to some (x, Θ) and satisfying $\Theta_n \in Q(x_n)$, we have $\Theta \in Q(x)$. Fix such a sequence and let $\varepsilon_n \to 0$. By the definition of $\varphi^o_-(x_n, \Theta_n)$, there exist $h_n \to 0+$, $x'_n \to x$ be such that for every n

$$\frac{\varphi(x'_n + h_n\Theta_n) - \varphi(x'_n)}{h_n} \leqq \varepsilon_n.$$

Consequently,

$$\varphi^o_-(x, \Theta) \leqq \liminf_{n \to \infty} \frac{\varphi(x'_n + h_n\Theta_n) - \varphi(x'_n)}{h_n} \leqq 0. \qquad \square$$

DEFINITION 3.6. Assume that φ is Lipschitz at $x_0 \in \Omega$. The *generalized gradient* of φ at x_0 is defined by

$$(18) \qquad \partial\varphi(x_0) = \{p \in \mathbf{R}^n \mid \forall \Theta \in \mathbf{R}^n, \ \varphi^o_-(x_0, \Theta) \leqq \langle p, \Theta \rangle\}.$$

We denote by $D^*\varphi(x_0)$ the set of all cluster points of gradients $\varphi'(x_n)$, when x_n converge to x_0 and φ is differentiable at x_n.

We note that

$$D^*\varphi(x_0) = \operatorname*{Lim\,sup}_{x \to x_0} \{\varphi'(x)\},$$

where $\mathrm{Lim\,sup}_{x \to x_0}$ denotes the upper limit when $x \to x_0$ (see [3, p. 41]).

In view of (17), the above definition of the generalized gradient is equivalent to the one given by Clarke. Comparing Definitions 3.1 and 3.6, we can easily realize that

$$(19) \qquad D^+\varphi(x_0) \subset \partial\varphi(x_0).$$

Moreover, it is clear that $D^*\varphi(x_0)$ is compact. From [11, Thm. 2.5.1], it follows that

$$(20) \qquad \partial\varphi(x_0) = \mathrm{co}\,(D^*\varphi(x_0)),$$

where co denotes the convex hull.

DEFINITION 3.7. Consider a convex subset K of $\mathbf{R}^n$. A function $\varphi : K \to \mathbf{R}$ is called semiconcave if there exists a function $\omega : \mathbf{R}_+ \times \mathbf{R}_+ \to \mathbf{R}_+$ such that

$$\begin{aligned} &\forall r \leqq R, \quad t \leqq T, \quad \omega(r, t) \leqq \omega(R, T), \quad \text{and} \\ &\forall R > 0, \quad \lim_{t \to 0+} \omega(R, t) = 0 \end{aligned} \tag{21}$$

and for every $R > 0$, $\lambda \in [0, 1]$ and any points $x, y \in K \cap RB$,

$$\lambda\varphi(x) + (1-\lambda)\varphi(y) \leqq \varphi(\lambda x + (1-\lambda)y) + \lambda(1-\lambda)\|x-y\|\omega(R, \|x-y\|).$$

We say that φ is semiconcave at x_0 if there exists a neighborhood of x_0 such that the restriction of φ to it is semiconcave. We call the above function ω a modulus of semiconcavity of φ.

Usually in the definition of semiconcavity $\omega(r, t) = ct$ for a nonnegative constant c (see [21], [22]), or $\omega(r, t) = ct^\alpha$ for $c \geqq 0$ and $\alpha \in]0, 1]$ [9]. We observe that every concave function $\varphi : K \to \mathbf{R}$ is semiconcave (with ω equal to zero). Furthermore, it is not difficult to check the following proposition.

PROPOSITION 3.8. *Let K be a convex subset of $\mathbf{R}^n$ and $\varphi : \mathbf{R}^n \to \mathbf{R}$ be continuously differentiable on a neighborhood of K. Then φ is semiconcave.*

Example 1. Consider a subset K of $\mathbf{R}^n$ and let $\operatorname{dist}(x, K)$ denote the distance from a point $x \in \mathbf{R}^n$ to K. Define the function $\varphi : \mathbf{R}^n \to \mathbf{R}_+$ by $\varphi(x) = \operatorname{dist}(x, K)^2$. Then, by a standard computation, φ is easily seen to be semiconcave.

In general, a Lipschitz function does not have directional derivatives. Our next aim is to show that for a semiconcave function, the directional derivatives exist and coincide with the regularized lower derivatives. This result was proved in [9], [10]. We provide a different proof of this fact for the sake of completeness.

THEOREM 3.9. *Let $x_0 \in \mathbf{R}^n$ and $\varphi : \mathbf{R}^n \to \mathbf{R}$ be Lipschitz and semiconcave at x_0. Then for every $\Theta \in \mathbf{R}^n$, the directional derivative $(\partial\varphi/\partial\Theta)(x_0)$ exists and is equal to the regularized lower derivative $\varphi^o_-(x_0, \Theta)$:*

$$\forall \Theta \in \mathbf{R}^n, \quad \frac{\partial\varphi}{\partial\Theta}(x_0) = \varphi^o_-(x_0, \Theta). \tag{22}$$

In particular, $D^+\varphi(x_0) \neq \emptyset$ and

$$D^+\varphi(x_0) = \partial\varphi(x_0) = \operatorname{co}(D^*\varphi(x_0)). \tag{23}$$

Consequently, the set-valued map $x \to D^+\varphi(x)$ is upper semicontinuous at x_0.

Proof. It is enough to consider the case $\|\Theta\| \leqq 1$. Let $\delta > 0$ be such that φ is semiconcave on $B_{2\delta}(x_0)$ with semiconcavity modulus $\omega(\cdot) := \omega(2\delta, \cdot)$. Fix $x \in B_\delta(x_0)$, $\Theta \in B$, and observe that for all $0 < h_1 \leqq h_2 \leqq \delta$ we have

$$\begin{aligned} \varphi(x + h_1\Theta) - \varphi(x) &= \varphi\left(\frac{h_1}{h_2}(x + h_2\Theta) + \left(1 - \frac{h_1}{h_2}\right)x\right) - \varphi(x) \\ &\geqq \frac{h_1}{h_2}\varphi(x + h_2\Theta) + \left(1 - \frac{h_1}{h_2}\right)\varphi(x) - \varphi(x) \\ &\quad - \frac{h_1}{h_2}\left(1 - \frac{h_1}{h_2}\right)h_2\|\Theta\|\omega(h_2\|\Theta\|) \\ &= \frac{h_1}{h_2}\varphi(x + h_2\Theta) - \frac{h_1}{h_2}\varphi(x) \\ &\quad - h_1\left(1 - \frac{h_1}{h_2}\right)\|\Theta\|\omega(h_2\|\Theta\|). \end{aligned}$$

Consequently, for all $0 < h_1 \leqq h_2 \leqq \delta$,

$$\frac{\varphi(x+h_1\Theta)-\varphi(x)}{h_1} \geqq \frac{\varphi(x+h_2\Theta)-\varphi(x)}{h_2} - \left(1-\frac{h_1}{h_2}\right)\omega(h_2\|\Theta\|)$$

and we proved that for every $x \in B_\delta(x_0)$

$$\forall 0 < h' \leqq h \leqq \delta,$$

$$\frac{\varphi(x+h'\Theta)-\varphi(x)}{h'} \geqq \frac{\varphi(x+h\Theta)-\varphi(x)}{h} - \omega(h\|\Theta\|). \tag{24}$$

Thus for every $0 < h \leqq \delta$,

$$\liminf_{h' \to 0+} \frac{\varphi(x_0+h'\Theta)-\varphi(x_0)}{h'} \geqq \frac{\varphi(x_0+h\Theta)-\varphi(x_0)}{h} - \omega(h\|\Theta\|).$$

Taking $\limsup_{h\to 0+}$ in the right-hand side of the above inequality yields that the directional derivative $(\partial\varphi/\partial\Theta)(x_0)$ does exist. Clearly, $(\partial\varphi/\partial\Theta)(x_0) \geqq \varphi_-^o(x_0,\Theta)$. To prove the opposite, fix $\varepsilon > 0$ and $0 < \lambda < \delta$. From the continuity of φ it follows that there exists $0 < \alpha < \delta$ such that for all $x \in B_\alpha(x_0)$,

$$\frac{\varphi(x_0+\lambda\Theta)-\varphi(x_0)}{\lambda} \leqq \frac{\varphi(x+\lambda\Theta)-\varphi(x)}{\lambda} + \varepsilon.$$

Thus, using (24), we obtain that

$$\frac{\varphi(x_0+\lambda\Theta)-\varphi(x_0)}{\lambda} \leqq \inf_{x\in B_\alpha(x_0), h\in]0,\lambda]} \frac{\varphi(x+h\Theta)-\varphi(x)}{h} + \omega(\lambda\|\Theta\|)+\varepsilon.$$

Letting ε, α, and λ converge to zero, we end the proof of the first statement. The second one results from (22) recalling (16), (18), and (20). □

PROPOSITION 3.10. *Let $\varphi : \mathbf{R}^n \to \mathbf{R}$ be Lipschitz and semiconcave at x_0. If $D^+\varphi(x_0)$ is a singleton, then φ is differentiable at x_0 and*

$$D^*\varphi(x_0) = \{\varphi'(x_0)\}.$$

In particular, if $D^+\varphi(x)$ is a singleton for all x near x_0, then φ is continuously differentiable at x_0.

The proof follows by exactly the same arguments as the ones in [9, Cor. 4.11 and 4.12].

DEFINITION 3.11. Let $K \subset \mathbf{R}^n$ be convex and $\varphi : K \to \mathbf{R}$ be given. Then φ is said to be semiconvex (respectively, semiconvex at x_0) whenever $-\varphi$ is semiconcave (respectively, semiconcave at x_0).

PROPOSITION 3.12. *Let $\varphi : \mathbf{R}^n \to \mathbf{R}$, $x_0 \in \mathbf{R}^n$. If φ is Lipschitz at x_0 and both semiconvex and semiconcave at x_0, then φ is continuously differentiable on a neighborhood of x_0.*

Proof. Since φ and $-\varphi$ are semiconcave at x_0 by Theorem 3.9, there exists a neighborhood $\mathcal{N}$ of x_0 such that for all $x \in \mathcal{N}$,

$$D^+\varphi(x) = \partial\varphi(x), \qquad D^+(-\varphi)(x) = \partial(-\varphi)(x).$$

Furthermore,

$$D^-\varphi(x) = -D^+(-\varphi)(x) = -\partial(-\varphi)(x) = \partial\varphi(x),$$

the last equality being a straightforward consequence of (20). Hence both $D^+\varphi(x)$ and $D^-\varphi(x)$ are nonempty and therefore φ is differentiable on $\mathcal{N}$. The conclusion follows from Proposition 3.10. □

4. Necessary and sufficient conditions for optimality. We begin this section with a sufficient condition for optimality which involves the superdifferential of the value function.

We associate with the control system (6) the Hamiltonian $H:[0,1]\times\mathbf{R}^n\times\mathbf{R}^n\to\mathbf{R}$ defined by

$$H(t,x,p)=\sup_{u\in U} < p, f(t,x,u)\rangle.$$

Under assumption (9), the function H is continuous, locally Lipschitz with respect to (x,p), and convex with respect to the third variable.

THEOREM 4.1. *Assume that* (9) *holds true and let* $\bar{x}:[0,1]\to\mathbf{R}^n$ *be a solution of the control system* (6), $\bar{x}(0)=\xi_0$ *and* $\bar{u}$ *be a corresponding control. If for almost every* $t\in[0,1]$, *there exists* $p(t)\in\mathbf{R}^n$ *such that*

$$(25)\qquad \begin{aligned}&\langle p(t),\bar{x}'(t)\rangle = H(t,\bar{x}(t),p(t)),\\ &(H(t,\bar{x}(t),p(t)),-p(t))\in D^+V(t,\bar{x}(t)),\end{aligned}$$

then $\bar{x}$ *is optimal for the problem* (7).

Proof. Consider the absolutely continuous function $\psi(t)=V(t,\bar{x}(t))$ and let $t\in[0,1]$ be such that the derivatives $\psi'(t)$ and $\bar{x}'(t)$ do exist and (25) holds true.

We first observe that (25) and (16) imply that

$$(26)\qquad \begin{aligned}0&=\langle((\langle p(t),\bar{x}'(t)\rangle,-p(t)),(1,\bar{x}'(t))\rangle\geqq\partial^+V(t,\bar{x}(t))(1,\bar{x}'(t))\\ &=\limsup_{h\to0+}\frac{V(t+h,\bar{x}(t)+h\bar{x}'(t))-V(t,\bar{x}(t))}{h}\\ &=\limsup_{h\to0+}\frac{V(t+h,\bar{x}(t+h))-V(t,\bar{x}(t))}{h}\\ &=\psi'(t).\end{aligned}$$

This yields that ψ is nonincreasing. Since the value function is also nondecreasing along trajectories of the control system (6), we deduce that the map $t\to V(t,\bar{x}(t))$ is constant. So $\bar{x}$ is optimal. □

The above map p may be constructed using the co-state variable of the maximum principle stated below.

THEOREM 4.2. *Assume that* (9) *holds true, that* f *is differentiable with respect to* x, *and* g *is differentiable. A trajectory control pair* $(\bar{x},\bar{u})$ *of the control system* (6) *with* $\bar{x}(0)=\xi_0$ *is optimal for the problem* (7) *if and only if the solution* $p:[0,1]\to\mathbf{R}^n$ *of the adjoint equation*

$$(27)\qquad -p'(t)=\left(\frac{\partial f}{\partial x}(t,\bar{x}(t),\bar{u}(t))\right)^*p(t),\, p(1)=-g'(\bar{x}(1))$$

satisfies the maximum principle

$$(28)\qquad \langle p(t),f(t,\bar{x}(t),\bar{u}(t))\rangle=\max_{u\in U}\langle p(t),f(t,\bar{x}(t),u)\rangle\quad \textit{a.e. in } [0,1]$$

and the generalized transversality conditions

$$(29)\qquad (H(t,\bar{x}(t),p(t)),-p(t))\in D^+V(t,\bar{x}(t))\quad \textit{a.e. in } [0,1],$$

$$(30)\qquad -p(t)\in D_x^+V(t,\bar{x}(t))\quad \forall t\in[0,1],$$

where $D_x^+V(t,\bar{x}(t))$ *denotes the superdifferential of* $V(t,\cdot)$ *at* $\bar{x}(t)$.

Furthermore, if V is semiconcave, then (29) *holds true everywhere in* $[0, 1]$.

We call such a function $p(\cdot)$ a co-state corresponding to the optimal trajectory $\bar{x}(\cdot)$.

Remark. The above theorem is a joint form of the maximum principle and the co-state inclusions (29), (30). In § 5 we provide a sufficient condition for V to be semiconcave.

Remark. The necessity of these conditions was proved in [17] under somewhat different assumptions. An inclusion on co-state p similar to (30) in the nonsmooth case was derived in [12]. An analogous statement is proved in [25] by a different method under stronger regularity assumptions. Necessary conditions in the form of generalized transversality conditions were proved in [26] as well. In this last paper, it is also shown that the subdifferentials of V and of $V(t, \cdot)$ along an optimal trajectory are at most singletons and are contained in the left-hand sides of (29) and (30), respectively. Another easy way to derive this fact is to use the remark following Definition 3.1.

Proof of Theorem 4.2. Sufficiency is a straightforward consequence of Theorem 4.1 and (28), (29). The fact that (27) and (28) are necessary is the well-known Pontryagin's maximum principle.

To prove the necessity of (29), fix $t \in [0, 1[$ such that $\bar{x}'(t) = f(t, \bar{x}(t), \bar{u}(t))$ and the equality (28) holds true, and let $\Theta \in \mathbf{R}^n$. Consider the solution $w(\cdot)$ of the linearized along $(\bar{x}, \bar{u})$ system

$$(31) \qquad \begin{aligned} &w'(s) = \frac{\partial f}{\partial x}(s, \bar{x}(s), \bar{u}(s))w(s), \quad s \in [t, 1], \\ &w(t) = \Theta. \end{aligned}$$

For every $h > 0$, let x_h be the solution to the differential equation

$$(32) \qquad \begin{aligned} &x'(s) = f(s, x(s), \bar{u}(s)), \quad s \in [t, 1], \\ &x(t) = \bar{x}(t) + h\Theta. \end{aligned}$$

From the variational equation we know that the quotients

$$\frac{x_h - \bar{x}}{h}$$

converge uniformly to w. Fix $\alpha \in \mathbf{R}$. Hence from (27) and (28), using that V is locally Lipschitz, nondecreasing along trajectories of (6), and constant along $\bar{x}$, we deduce that

$$\begin{aligned} &\partial^+ V(t, \bar{x}(t))(\alpha, \alpha\bar{x}'(t) + \Theta) \\ &\quad = \limsup_{h \to 0+} (V(t + \alpha h, \bar{x}(t) + h(\alpha\bar{x}'(t) + w(t))) - V(t, \bar{x}(t)))/h \\ &\quad = \limsup_{h \to 0+} (V(t + \alpha h, \bar{x}(t + \alpha h) + hw(t + \alpha h)) - V(t, \bar{x}(t)))/h \\ &\quad = \limsup_{h \to 0+} (V(t + \alpha h, x_h(t + \alpha h)) - V(t, \bar{x}(t)))/h \\ &\quad \leqq \limsup_{h \to 0+} (g(x_h(1)) - g(\bar{x}(1)))/h \\ &\quad = \langle g'(\bar{x}(1)), w(1)\rangle = \langle -p(t), w(t)\rangle \\ &\quad = \langle -p(t), -\alpha\bar{x}'(t)\rangle + \langle -p(t), \alpha\bar{x}'(t) + \Theta\rangle \\ &\quad = \alpha H(t, \bar{x}(t), p(t)) + \langle -p(t), \alpha\bar{x}'(t) + \Theta\rangle. \end{aligned}$$

Hence we deduce that for every $\alpha \in \mathbf{R}$ and $\Theta_1 \in \mathbf{R}^n$,

$$\partial^+ V(t, \bar{x}(t))(\alpha, \Theta_1) \leqq \alpha H(t, \bar{x}(t), p(t)) + \langle -p(t), \Theta_1 \rangle$$

and the proof of (29) follows from (16). To prove (30), observe that for every $t \in [0, 1]$, $\Theta \in \mathbf{R}^n$, and the solution w of (31),

$$\begin{aligned}\langle -p(t), \Theta \rangle &= \langle g'(\bar{x}(1)), w(1) \rangle \\ &\geqq \limsup_{h \to 0+} \left(V(t, \bar{x}(t) + h\Theta) - V(t, \bar{x}(t)) \right)/h = \partial_x^+ V(t, \bar{x}(t))(\Theta).\end{aligned}$$

This and (16) imply (30). When V is semiconcave, then the last statement follows from (29), (23), the continuity of $H(\cdot)$, $p(\cdot)$, $\bar{x}(\cdot)$, and Theorem 3.9. □

Remark. It can be shown that the conclusion of Theorem 4.2 remains true even when g is not differentiable, but $D^+g(\bar{x}(1) \neq \emptyset$. In this case we can take for $p(1)$ any element of $D^+g(\bar{x}(1))$. The same observation applies to Theorem 4.5 below.

Remark. For systems governed by an autonomous state equation—that is, with f independent of t—we can easily show that the Hamiltonian is constant along any optimal trajectory/co-state pair $(\bar{x}, p)$. Indeed, recalling (28), (27), for almost all $t, s \in [0, 1]$, we obtain

$$\begin{aligned}H(\bar{x}(t), p(t)) - H(\bar{x}(s), p(s)) &\geqq \langle p(t), f(\bar{x}(t), \bar{u}(s)) - f(\bar{x}(s), \bar{u}(s)) \rangle \\ &\quad + \langle p(t) - p(s), f(\bar{x}(s), \bar{u}(s)) \rangle = o(|t-s|).\end{aligned}$$

Since the above argument is symmetric,

$$|H(\bar{x}(t), p(t)) - H(\bar{x}(s), p(s))| \leqq o(|t-s|).$$

Since $t \to H(\bar{x}(t), p(t))$ is absolutely continuous, the above inequality implies that $H(\bar{x}(t), p(t))$ is constant.

When the Hamiltonian H is differentiable with respect to (x, p) at $(\bar{x}(t), p(t))$ for all $t \in [0, 1]$, then $\bar{x}$ and the co-state p satisfy the Hamiltonian system

$$\bar{x}'(t) = \frac{\partial H}{\partial p}(t, \bar{x}(t), p(t)),$$

$$p'(t) = -\frac{\partial H}{\partial x}(t, \bar{x}(t), p(t)).$$

More generally arguments similar to [19, Remark 4.10] imply the following proposition.

PROPOSITION 4.3. *Let $(t, \bar{x}, \bar{p}) \in [0, 1] \times \mathbf{R}^n \times \mathbf{R}^n$ and $\bar{u} \in U$ be such that*

$$\langle p, f(t, \bar{x}, \bar{u}) \rangle = H(t, \bar{x}, \bar{p}).$$

(i) *If $H(t, \cdot, \bar{p})$ is differentiable at $\bar{x}$, then*

$$\frac{\partial H}{\partial x}(t, \bar{x}, \bar{p}) = \left(\frac{\partial f}{\partial x}(t, \bar{x}, \bar{u}) \right)^* \bar{p}.$$

(ii) *If $H(t, \bar{x}, \cdot)$ is differentiable at $\bar{p}$, then*

$$\frac{\partial H}{\partial p}(t, \bar{x}, \bar{p}) = f(t, \bar{x}, \bar{u}).$$

It is well known that for every $(t, x) \in [0, 1] \times \mathbf{R}^n$ at which V is differentiable, we have

$$-\frac{\partial V}{\partial t}(t, x) + H\left(t, x, -\frac{\partial V}{\partial x}(t, x) \right) = 0. \tag{33}$$

When V is not differentiable at (t, x), the above equation has to be understood in the viscosity sense (see [13], [14]).

Since the Hamiltonian is continuous, we immediately deduce from (33) that

$$\text{(34)} \qquad \forall (t,x) \in [0,1]\times \mathbf{R}^n, \quad \forall (p_t, p_x) \in D^* V(t,x), \quad -p_t + H(t,x,-p_x) = 0.$$

Moreover, since $H(t, x, \cdot)$ is convex, (20) yields that

$$\text{(35)} \qquad \forall (t,x) \in]0,1[\times \mathbf{R}^n, \quad \forall (p_t, p_x) \in \partial V(t,x), \quad -p_t + H(t,x,-p_x) \leqq 0.$$

In particular, as $D^+ V(t,x) \subset \partial V(t,x)$, the inequality in (35) holds for all $(p_t, p_x) \in D^+ V(t,x)$ (in fact, by [14], [15], this inequality is the definition of viscosity subsolution).

The following is an adaptation of [26, Prop. 3.1].

PROPOSITION 4.4. *Assume* (9) *and let* $\bar{x}$ *be an optimal solution of problem* (8). *Then for every* $t \in]t_0, 1[$

$$\text{(36)} \qquad \forall (p_t, p_x) \in D^+ V(t, \bar{x}(t)), \quad -p_t + H(t, \bar{x}(t), -p_x) = 0.$$

Consequently, if for every $t \in]t_0, 1[$, $H(t, \bar{x}(t), \cdot)$ *is strictly convex, then* $D^+ V(t, \bar{x}(t))$ *is at most a singleton for every* $t \in]t_0, 1[$.

Proof. Let $\bar{u}$ be an optimal control corresponding to $\bar{x}$ and let $t \in]t_0, 1[$. Consider

$$\text{(37)} \qquad v \in \overline{\text{co}}\, f(t, \bar{x}(t), U)$$

such that for some $h_n \to 0+$, $\lim_{n\to\infty} (\bar{x}(t-h_n) - \bar{x}(t))/h_n = -v$. Then for all $(p_t, p_x) \in D^+ V(t, \bar{x}(t))$,

$$0 \geqq \limsup_{s \to t-} \frac{V(s, \bar{x}(s)) - V(t, \bar{x}(t)) - p_t(s-t) - \langle p_x, \bar{x}(s) - \bar{x}(t)\rangle}{|s-t| + \|\bar{x}(s) - \bar{x}(t)\|}.$$

Since $V(\cdot, \bar{x}(\cdot))$ is constant, the above estimate yields

$$0 \geqq \lim_{h \to \infty} \frac{-p_t(-h_n) - \langle p_x, \bar{x}(t-h_n) - \bar{x}(t)\rangle}{h_n} = p_t - \langle p_x, -v\rangle.$$

So we derived that

$$0 \leqq -p_t + \langle -p_x, v\rangle \leqq -p_t + H(t, x, -p_x)$$

From this last inequality and (35), we obtain (36). If $H(t, \bar{x}(t), \cdot)$ is strictly convex, then for all $(p_t^i, p_x^i) \in D^+ V(t, \bar{x}(t))$, $i = 1, 2$,

$$\begin{aligned} \tfrac{1}{2}(p_t^1 + p_t^2) &= \tfrac{1}{2}(H(t, \bar{x}(t), -p_x^1) + H(t, \bar{x}(t), -p_x^2)) \\ &\leqq H(t, \bar{x}(t), -\tfrac{1}{2}(p_x^1 + p_x^2)) = \tfrac{1}{2}(p_t^1 + p_t^2). \end{aligned}$$

If $(p_t^1, p_x^1) \neq (p_t^2, p_x^2)$, then in the above we will get a strict inequality, which would lead to a contradiction. Therefore $(p_t^1, p_x^1) = (p_t^2, p_x^2)$. □

We show next that, whenever $p(0) = -V_x'(0, \xi_0)$, we have the equality in the inclusion (30).

THEOREM 4.5. *Assume that* (9) *holds true, that* f *is differentiable with respect to* x *and* g *is differentiable. Suppose, furthermore, that the derivative* $V_x'(t_0, x_0)$ *exists and let* $\bar{x}$ *be an optimal solution for the problem* (8). *Consider the co-state* $p : [t_0, 1] \to \mathbf{R}^n$ *corresponding to* $\bar{x}$ *and given by Theorem* 4.2, *where the interval* $[0, 1]$ *is replaced by* $[t_0, 1]$ *and* ξ_0 *by* x_0. *Then*

$$\{-p(t)\} = D_x^+ V(t, \bar{x}(t)) \quad \forall t \in [t_0, 1].$$

In the next section, we show that under some additional regularity assumptions on f, $p(t)$ is equal to the derivative of the value function $V'_x(t, \bar{x}(t))$ for all t, provided $V'_x(t_0, x_0)$ exists.

Proof. We already know from Theorem 4.2 that

$$-p(t) \in D_x^+ V(t, \bar{x}(t)) \quad \forall t \in [t_0, 1].$$

Thus $p(t_0) = -V'_x(t_0, x_0)$.

Let $\bar{u}$ be an optimal control corresponding to $\bar{x}$. Fix Θ and let w, x_h have the same meaning as in the proof of Theorem 4.2 with t replaced by t_0. Then, since V is nondecreasing along trajectories of the control system (6) and constant along $\bar{x}$, for all $t \in [t_0, 1]$,

$$\begin{aligned}\langle p(t_0), \Theta\rangle = \langle -V'_x(t_0, x_0), \Theta\rangle &= -\lim_{h\to 0+} \frac{V(t_0, x_0 + h\Theta) - V(t_0, x_0)}{h} \\ &\geqq -\limsup_{h\to 0+} \frac{V(t, x_h(t)) - V(t, \bar{x}(t))}{h} \\ &= -\limsup_{h\to 0+} \frac{V(t, \bar{x}(t) + hw(t)) - V(t, \bar{x}(t))}{h} = -\partial_x^+ V(t, \bar{x}(t))(w(t)),\end{aligned}$$

where $\partial_x^+ V(t, \bar{x}(t))(w(t))$ denotes the upper Dini derivative of $V(t, \cdot)$ at $\bar{x}(t)$ in the direction $w(t)$.

Using (16) we deduce that for every $q \in D_x^+ V(t, \bar{x}(t))$, we have

$$\langle p(t_0), \Theta\rangle \geqq \langle -q, w(t)\rangle = \langle -q, X(t)\Theta\rangle = \langle -X(t)^* q, \Theta\rangle,$$

where X denotes the fundamental solution of

$$\begin{aligned}&X'(t) = \frac{\partial f}{\partial x}(t, \bar{x}(t), \bar{u}(t))X(t), \qquad t \in [t_0, 1],\\ &X(t_0) = Id.\end{aligned}$$

Since $\Theta \in \mathbf{R}^n$ is arbitrary, we have $p(t_0) = -X(t)^* q$. On the other hand, $p(\cdot)$ being a solution of (27), we know that $p(t_0) = X(t)^* p(t)$. Since for every $t \in [t_0, 1]$, the matrix $X(t)$ is nonsingular, we have proved that $-p(t) = q$. This yields that $D_x^+ V(t, \bar{x}(t))$ is single-valued and ends the proof. □

Whenever H happens to be more regular, we can prove the following theorem concerning optimal design. For every (t_0, x_0), define

$$D_x^* V(t_0, x_0) = D^* W(x_0), \tag{38}$$

where W is given by $W(x) = V(t_0, x)$.

THEOREM 4.6. *Assume that* (9) *holds true, that f is differentiable with respect to x, g is differentiable, and that $H(t, \cdot, \cdot)$ is differentiable on $\mathbf{R}^n \times (\mathbf{R}^n \backslash \{0\})$ for almost every $t \in [0, 1]$. Furthermore, assume that the sets $f(t, x, U)$ are convex and compact and for every $R > 1$ there exists a nonnegative integrable function $l_R \in L^1(0, 1; \mathbf{R}_+)$ such that for all $x, y \in RB$ and $p, q \in RB \backslash (1/R)B$*

$$\begin{aligned}&\left\|\frac{\partial H}{\partial x}(t, x, p) - \frac{\partial H}{\partial x}(t, y, q)\right\| + \left\|\frac{\partial H}{\partial p}(t, x, p) - \frac{\partial H}{\partial p}(t, y, q)\right\| \\ &\qquad \leqq l_R(t)(\|x - y\| + \|p - q\|).\end{aligned} \tag{39}$$

Let $(t_0, x_0) \in [0, 1] \times \mathbf{R}^n$ *and* $p_0 \neq 0$ *be such that*

$$-p_0 \in D_x^* V(t_0, x_0). \tag{40}$$

Then the Hamiltonian system

$$\begin{aligned} &x'(t) = \frac{\partial H}{\partial p}(t, x(t), p(t)), \qquad x(t_0) = x_0, \\ &p'(t) = -\frac{\partial H}{\partial x}(t, x(t), p(t)), \qquad p(t_0) = p_0, \\ &p(t) \neq 0, \qquad t \in [t_0, 1] \end{aligned} \tag{41}$$

has a unique solution $(\bar{x}(\cdot), \bar{p}(\cdot))$ *defined on* $[t_0, 1]$. *Moreover,* $\bar{x}(\cdot)$ *is an optimal solution of problem* (8).

Furthermore, if $g'(\cdot)$ *is continuous at* $\bar{x}(1)$, *then* $\bar{p}(\cdot)$ *is the co-state corresponding to* $\bar{x}(\cdot)$.

Remark. The above theorem extends a result of [10], which concerned a problem in calculus of variations. For such problems, condition (39) is natural. It is much more restrictive for nonlinear control systems. It is well known that $H(t, x, \cdot)$ is not differentiable at zero when $f(t, x, U)$ is not a singleton. This is why we eliminate $p = 0$ in our assumptions.

We observe that (39) is satisfied whenever the variables x and u are "separated":

$$f(t, x, u) = \varphi(t, x) + \psi(t, u),$$

where $\varphi(t, \cdot)$ has $k_R(t)$ Lipschitz gradient on RB for some $k_R \in L^1(0, 1; \mathbf{R}_+)$ and the boundary of $\psi(t, U)$ is sufficiently smooth.

Proof. By the very definition of $D_x^* V(t_0, x_0)$, it follows that there exists a sequence x_k converging to x_0 such that $V(t_0, \cdot)$ is differentiable at x_k and

$$-p_0 = \lim_{k \to \infty} \frac{\partial V}{\partial x}(t_0, x_k).$$

Let $(\bar{x}_k, u_k)$ be an optimal trajectory control pair for problem (8) with x_0 replaced by x_k. By Theorem 4.2 (applied with the interval $[0, 1]$ replaced by $[t_0, 1]$ and ξ_0 by x_k) there exists an absolutely continuous function $\bar{p}_k : [t_0, 1] \to \mathbf{R}^n$ such that

$$\begin{aligned} &-\bar{p}_k'(t) = \left(\frac{\partial f}{\partial x}(t, \bar{x}_k(t), \bar{u}_k(t))\right)^* \bar{p}_k(t) \quad \text{a.e. in } [t_0, 1], \\ &-\bar{p}_k(t_0) = \frac{\partial V}{\partial x}(t_0, x_k). \end{aligned} \tag{42}$$

Therefore, $p_k(t) \neq 0$ for all $t \in [t_0, 1]$ for sufficiently large k. By Proposition 4.3, for almost every $t \in [t_0, 1]$,

$$\begin{aligned} &\bar{x}_k'(t) = \frac{\partial H}{\partial p}(t, x_k(t), \bar{p}_k(t)), \qquad \bar{x}_k(t_0) = x_k, \\ &\bar{p}_k'(t) = -\frac{\partial H}{\partial x}(t, \bar{x}_k(t), \bar{p}_k(t)), \\ &\bar{p}_k(t_0) = -\frac{\partial V}{\partial x}(t_0, x_k). \end{aligned} \tag{43}$$

Recalling assumption (9) (iii), we conclude that $\bar{x}_k$, $k=1,\cdots$ are equicontinuous and equibounded. Furthermore, from (9)(ii) and (42) it follows that $\bar{p}_k$ are also equicontinuous and equibounded and the maps $t\to(\partial f/\partial x)(t,\bar{x}_k(t),u_k(t))$ are integrably bounded on $[t_0,1]$. So, taking a subsequence and keeping the same notation, we may assume that $(\bar{x}_k,\bar{p}_k)$ converge uniformly to some $(\bar{x},\bar{p})$ and $(\partial f/\partial x)(\cdot,\bar{x}_k(\cdot),u_k(\cdot))$ converge weakly in $L^1(t_0,1)$ to some $A(\cdot)$. In particular $\bar{p}(t_0)=p_0\neq 0$ and $\bar{p}$ solves the linear equation

$$-\bar{p}'(t)=A(t)^*\bar{p}(t)\quad \text{a.e. in } [t_0,1].$$

Thus $\bar{p}(t)\neq 0$ for all $t\in[t_0,1]$. Fix $R>1$ so that

$$\forall s\in[t_0,1],\quad \frac{2}{R}\leqq\|\bar{p}(s)\|\leqq\frac{R}{2}.$$

Then, for all sufficiently large k and all $s\in[t_0,1]$, we have $1/R\leqq\|p_k(s)\|\leqq R$. So, using (39) and taking the limit in (43), we deduce $(\bar{x},\bar{p})$ is a solution of the Hamiltonian system (41).

Since $\bar{p}$ never vanishes, the assumption (39) implies that $(\bar{x},\bar{p})$ is the only solution of (41). On the other hand,

$$V(t_0,x_0)=\lim_{k\to\infty}V(t_0,\bar{x}_k(t_0))=\lim_{k\to\infty}g(\bar{x}_k(1))=g(\bar{x}(1))$$

and therefore $\bar{x}$ is optimal for the problem (8).

Assume next that g' is continuous at $\bar{x}(1)$. Then $\bar{p}(1)=-g'(\bar{x}(1))\neq 0$. Let p_1 be a co-state corresponding to the optimal trajectory $\bar{x}$. Then $p_1(t)\neq 0$ for all $t\in[t_0,1]$ and, by Proposition 4.3, it solves the equation

$$(44)\qquad -p'(t)=\frac{\partial H}{\partial x}(t,\bar{x}(t),p(t))\quad \text{a.e. in } [t_0,1],\quad p(1)=-g'(\bar{x}(1)).$$

Since $\bar{p}$ is also a solution of (44), we deduce that $p_1=\bar{p}$ by uniqueness. The proof is complete. □

Remark. (i) By minor modifications of the above arguments it is easy to show that condition (40) may be replaced by the following one:

$$(45)\qquad (H(t_0,x_0,p_0),-p_0)\in D^*V(t_0,x_0).$$

In general, (40) and (45) are not comparable. If V is semiconcave, however, then (40) is more restrictive than (45). This can be shown using Proposition 5.2 below.

(ii) In calculus of variations, the necessity of (45) was proved in [10].

Other examples of problems for which (40) is necessary are given by optimal control problems having a unique optimal trajectory for the initial state (t_0,x_0).

THEOREM 4.7. *Assume* (9), *that g is continuously differentiable, that f is differentiable with respect to x, that* $f(t,x,U)$ *are convex and compact, and that H is continuously differentiable with respect to x. Further assume that, for every* $R>0$, *there exists a nonnegative integrable function* $l_R\in L^1(0,1;\mathbf{R}_+)$ *such that*

$$(46)\qquad \forall x,y,p\in RB,\quad \left\|\frac{\partial H}{\partial x}(t,x,p)-\frac{\partial H}{\partial x}(t,y,p)\right\|\leqq l_R(t)\|x-y\|.$$

If the problem (8) *has a unique optimal solution* $\bar{x}$, *then* $V(t,\cdot)$ *is differentiable at* $\bar{x}(t)$ *for all* $t\in[t_0,1]$.

Proof. Observe that for every $t\in[t_0,1]$, the problem (8) has a unique optimal solution for (t_0,x_0) replaced by $(t,\bar{x}(t))$. For this reason we only check that $V(t_0,\cdot)$ is differentiable at x_0.

By [11, pp. 33, 63] it suffices to show that $D_x^*V(t_0,x_0)$ is a singleton. Let $p_1,p_2\in D_x^*V(t_0,x_0)$ and consider sequences $\{x_k^1\}$ and $\{x_k^2\}$ converging to x_0, such that

$$\lim_{k\to+\infty}\frac{\partial V}{\partial x}(t_0,x_k^i)=p_i,\qquad i=1,2.$$

Let $\bar{x}_k^i$ be an optimal trajectory for problem (8) with x_0 replaced by x_k^i, $i=1,2$ and denote by p_k^i a corresponding co-state. Then, by Proposition 4.3,

$$\begin{aligned}(p_k^i)'(t)&=-\frac{\partial H}{\partial x}(t,\bar{x}_k^i(t),p_k^i(t)),\\ p_k^i(1)&=-g'(\bar{x}_k^i(1)),\\ p_k^i(t_0)&=-\frac{\partial V}{\partial x}(t_0,x_k^i).\end{aligned}$$

By assumption (9)(iii), $\bar{x}_k^i$ are bounded and equicontinuous. Since the solution of (8) is unique, we deduce that $\bar{x}_k^i$ converge uniformly to $\bar{x}$ for $i=2$. Taking subsequences and keeping the same notation, we may assume that p_k^i converge uniformly to the unique solution p of

$$p'(t)=-\frac{\partial H}{\partial x}(t,\bar{x}(t),p(t)),\qquad p(1)=-g'(\bar{x}(1)).$$

In particular, $p_1=p(t_0)=p_2$ and so $D_x^*V(t_0,x_0)$ is a singleton. □

THEOREM 4.8. *Under the assumptions of Theorem* 4.6, *assume that g is continuously differentiable. Then* $V(t_0,\cdot)$ *is differentiable at* x_0 *with derivative different from zero if and only if there is a unique optimal trajectory* $\bar{x}$ *for problem* (8) *and* $g'(\bar{x}(1))\neq 0$.

Proof. Assume that $(\partial V/\partial x)(t_0,x_0)\neq 0$. Let $\bar{x}$ be optimal for problem (8). By Theorem 4.2, $g'(\bar{x}(1))\neq 0$. By Proposition 4.3, every optimal trajectory/co-state pair solves the Hamiltonian system (41). This yields the uniqueness of optimal trajectory.

Conversely, assume that (8) has a unique optimal solution $\bar{x}$ and $g'(\bar{x}(1))\neq 0$. By Theorem 4.7, $V(t_0,\cdot)$ is differentiable at x_0. Theorem 4.2 implies that $(\partial V/\partial x)(t_0,x_0)\neq 0$. □

5. Semiconcavity properties of the value function.

We provide a sufficient condition for the semiconcavity of the value function $V:[0,1]\times\mathbf{R}^n\to\mathbf{R}$ introduced in the first section. Throughout the whole section we assume the following:

(47)
- (i) $f:[0,1]\times\mathbf{R}^n\times U\to\mathbf{R}^n$ is continuous,
- (ii) $\exists M>0$ such that $\forall(t,x,u)\in[0,1]\times\mathbf{R}^n\times U$, $\|f(t,x,u)\|\leqq M(\|x\|+1)$,
- (iii) $\exists L>0$ such that $\forall t_1,t_2\in[0,1],\forall x_1,x_2\in\mathbf{R}^n,\forall u\in U$, $\|f(t_1,x_1,u)-f(t_2,x_2,u)\|\leqq L(|t_1-t_2|+\|x_1-x_2\|)$,
- (iv) $\exists\omega:\mathbf{R}_+\times\mathbf{R}_+\to\mathbf{R}_+$ such that (21) holds true and $\forall\lambda\in[0,1],\forall u\in U,\forall R>0,\forall x_0,x_1\in RB,\forall t\in[0,1]$, $\|\lambda f(t,x_0,u)+(1-\lambda)f(t,x_1,u)-f(t,\lambda x_0+(1-\lambda)x_1,u)\|$ $\leqq\lambda(1-\lambda)\|x_1-x_0\|\omega(R,\|x_1-x_0\|)$,
- (v) $g:\mathbf{R}^n\to\mathbf{R}$ is locally Lipschitz and semiconcave.

Remark. (1) Assumption (47)(iv) holds true in particular when f is continuously differentiable with respect to x uniformly in (t,u). More precisely, (47)(iv) is satisfied if we assume that there exists a function $\omega: \mathbf{R}_+ \times \mathbf{R}_+ \to \mathbf{R}_+$ satisfying (21) such that

$$\left\| \frac{\partial f}{\partial x}(t, x_1, u) - \frac{\partial f}{\partial x}(t, x_2, u) \right\| \leqq \omega(R, \|x_1 - x_2\|)$$

for all $u \in U$, $t \in [0, 1]$, $x_1, x_2 \in RB$.

(2) Vice versa, Proposition 3.12 implies that, if f satisfies (47)(iv), then f is continuously differentiable with respect to x.

The main result of this section is the following theorem.

THEOREM 5.1. *If* (47) *holds true, then the value function is semiconcave on* $[0, 1] \times \mathbf{R}^n$.

Proof. For every $t \in [0, 1]$ and measurable function $u: [t, 1] \to U$, we denote by $y(\cdot\,; t, x, u)$ the solution of the system

$$y'(s) = f(s, y(s), u(s)), \qquad y(t) = x.$$

The Gronwall lemma implies that

$$\forall x \in RB, \quad \forall s \in [t, 1], \quad \|y(s)\| \leqq C_R := (R + M)\, e^M. \tag{48}$$

Moreover, for all $t \in [0, 1]$, $s \in [t, 1]$, $x_0, x_1 \in \mathbf{R}^n$ and all measurable functions $u: [t, 1] \to U$, we have

$$\|y(s; t, x_1, u) - y(s; t, x_0, u)\| \leqq e^{L(s-t)} \|x_1 - x_0\|. \tag{49}$$

Step 1. We claim that there exists $\omega_1: \mathbf{R}_+ \times \mathbf{R}_+ \to \mathbf{R}_+$ satisfying (21) such that for all $0 \leqq t \leqq s \leqq 1$, $R > 0$, $x_0, x_1 \in RB$, $\lambda \in [0, 1]$, and measurable functions $u: [t, 1] \to U$, we have

$$\begin{aligned} &\|\lambda y(s; t, x_1, u) + (1 - \lambda) y(s; t, x_0, u) - y(s; t, \lambda x_0 + (1 - \lambda) x_1, u)\| \\ &\quad \leqq \lambda(1 - \lambda) \|x_1 - x_0\| \omega_1(R, \|x_1 - x_0\|). \end{aligned}$$

Indeed, set $x_\lambda = \lambda x_0 + (1 - \lambda) x_1$ and define

$$y_\lambda(\tau) = \lambda y(\tau; t, x_1, u) + (1 - \lambda) y(\tau; t, x_0, u) - y(\tau; t, x_\lambda, u).$$

Then

$$\begin{aligned} y'_\lambda(\tau) &= \lambda f(\tau, y(\tau; t, x_1, u), u(\tau)) + (1 - \lambda) f(\tau, y(\tau; t, x_0, u), u(\tau)) \\ &\qquad - f(\tau, y(\tau; t, x_\lambda, u), u(\tau)), \\ y_\lambda(t) &= 0. \end{aligned}$$

Thus by assumptions (47)

$$\begin{aligned} \|y'_\lambda(\tau)\| \leqq \lambda(1 - \lambda) &\|y(\tau; t, x_1, u) - y(\tau; t, x_0, u)\| \\ &\cdot \omega(C_R, \|y(\tau; t, x_1, u) - y(\tau; t, x_0, u)\|) + L\|y_\lambda(\tau)\|, \end{aligned}$$

where C_R is defined in (48). Our claim follows from (49) and the Gronwall lemma.

Step 2. We claim that there exists $\omega_2: \mathbf{R}_+ \times \mathbf{R}_+ \to \mathbf{R}_+$ satisfying (21) such that for all $t \in [0, 1]$, $\lambda \in [0, 1]$, $R > 0$, and $x_0, x_1 \in RB$ the following inequality holds true:

$$\begin{aligned} &\lambda V(t, x_1) + (1 - \lambda) V(t, x_0) - V(t, \lambda x_1 + (1 - \lambda) x_0) \\ &\quad \leqq \lambda(1 - \lambda) \|x_1 - x_0\| \omega_2(R, \|x_1 - x_0\|). \end{aligned}$$

Indeed, define x_λ as in Step 1, fix $\varepsilon > 0$ and a control u_ε such that

$$V(t, x_\lambda) > g(y(1; t, x_\lambda, u_\varepsilon)) - \varepsilon.$$

Let ω_g denote a modulus of semiconcavity of g and L_R a Lipschitz constant of g on the ball of radius C_R. Then from (49) and Step 1, we get

$$\begin{aligned}\lambda V(t,x_1)&+(1-\lambda)V(t,x_0)-V(t,x_\lambda)\\&<\lambda g(y(1;t,x_1,u_\varepsilon))+(1-\lambda)g(y(1;t,x_0,u_\varepsilon))-g(y(1;t,x_\lambda,u_\varepsilon))+\varepsilon\\&\leqq \lambda(1-\lambda)\|y(1;t,x_1,u_\varepsilon)-y(1;t,x_0,u_\varepsilon)\|\\&\quad\cdot\omega_g(C_R,\|y(1;t,x_1,u_\varepsilon)-y(1;t,x_0,u_\varepsilon)\|)\\&\quad+L_R\|\lambda y(1;t,x_1,u_\varepsilon)+(1-\lambda)y(1;t,x_0,u_\varepsilon)-y(1;t,x_\lambda,u_\varepsilon)\|+\varepsilon\\&\leqq L'_R\lambda(1-\lambda)\|x_1-x_0\|\\&\quad\cdot(\omega_g(C_R,e^L\|x_1-x_0\|)+\omega_1(R,\|x_1-x_0\|))+\varepsilon\end{aligned}$$

for some L'_R depending only on L_R and L. Since $\varepsilon>0$ is arbitrary, our claim follows. Thus we proved the semiconcavity of $V(t,\cdot)$.

Step 3. Consider next $0\leqq t_1<t_0\leqq 1$, $R>0$, and let $x_0,x_1\in RB$, $\lambda\in[0,1]$. Define

$$x_\lambda=\lambda x_1+(1-\lambda)x_0,\qquad t_\lambda=\lambda t_1+(1-\lambda)t_0.$$

Pick any $\varepsilon>0$ and let u_ε be such that

$$V(t_0,y(t_0;t_\lambda,x_\lambda,u_\varepsilon))<V(t_\lambda,x_\lambda)+\varepsilon.$$

Define

$$\tau(s)=\begin{cases}\lambda s+(1-\lambda)t_0, & \text{if } t_1\leqq s\leqq t_0,\\ s & \text{otherwise.}\end{cases}\tag{50}$$

Since the value function is nondecreasing along trajectories of our control system, we have

$$\begin{aligned}\lambda V(t_1,x_1)&+(1-\lambda)V(t_0,x_0)-V(t_\lambda,x_\lambda)\\&\leqq\lambda V(t_0,y(t_0;t_1,x_1,u_\varepsilon\circ\tau))+(1-\lambda)V(t_0,x_0)\\&\quad-V(t_0,y(t_0;t_\lambda,x_\lambda,u_\varepsilon))+\varepsilon.\end{aligned}\tag{51}$$

Set $y_1(s)=y(s;t_1,x_1,u_\varepsilon\circ\tau)$, $y_\lambda(s)=y(s;t_\lambda,x_\lambda,u_\varepsilon)$, and let K_R denote the Lipschitz constant of V on $[0,1]\times C_RB$. By (51) and Step 2 we obtain

$$\begin{aligned}\lambda V(t_1,x_1)&+(1-\lambda)V(t_0,x_0)-V(t_\lambda,x_\lambda)\\&\leqq\lambda(1-\lambda)\|y_1(t_0)-x_0\|\omega_2(C_R,\|y_1(t_0)-x_0\|)\\&\quad+K_R\|\lambda y_1(t_0)+(1-\lambda)x_0-y_\lambda(t_0)\|.\end{aligned}\tag{52}$$

On the other hand, from assumption (47)(ii) it follows that

$$\forall s\in[t_1,t_0],\quad \|y_1(s)-x_0\|\leqq\|x_1-x_0\|+M_R(t_0-t_1),\tag{53}$$

where $M_R=M(1+C_R)$. Set

$$z(s)=\lambda y_1(\tau^{-1}(s))+(1-\lambda)x_0-y_\lambda(s)$$

and note that $z(t_\lambda)=0$, $z(t_0)=\lambda y_1(t_0)+(1-\lambda)x_0-y_\lambda(t_0)$. Furthermore, using (47)(iii), we obtain the following estimates:

$$\begin{aligned}\|z'(s)\|&=\|f(\tau^{-1}(s),y_1\circ\tau^{-1}(s),u_\varepsilon(s))-f(s,y_\lambda(s),u_\varepsilon(s))\|\\&\leqq L(|\tau^{-1}(s)-s|+\|y_1\circ\tau^{-1}(s)-y_\lambda(s)\|)\\&\leqq L\|z(s)\|+L(1-\lambda)\|y_1\circ\tau^{-1}(s)-x_0\|+L\frac{1-\lambda}{\lambda}(t_0-s).\end{aligned}$$

Therefore from the Gronwall inequality and (53) we deduce that

$$\begin{aligned}\|z(t_0)\| &\leqq L(1-\lambda)\int_{t_\lambda}^{t_0}\left(\|y_1\circ\tau^{-1}(s)-x_0\|+\frac{1}{\lambda}(t_0-s)\right)e^{L(t_0-s)}\,ds\\ &\leqq L\,e^L\lambda(1-\lambda)(t_0-t_1)\left(\|x_1-x_0\|+\left(\frac{1}{2}+M_R\right)(t_0-t_1)\right).\end{aligned}\tag{54}$$

Inequalities (52), (53), and (54) imply the conclusion. □

PROPOSITION 5.2. *Assume that the value function is semiconcave at a point* $(t_0, x_0)\in[0,1]\times\mathbf{R}^n$. *If* $D_x^+V(t_0,x_0)$ *is a singleton, then* V *is differentiable at* (t_0,x_0) *and* $D^*V(t_0,x_0)=\{V'(t_0,x_0)\}$.

Here at boundary points ($t_0\in\{0,1\}$), the above differentiability, of course, has to be understood in the one-sided sense.

Proof. Let $\pi_x:\mathbf{R}\times\mathbf{R}^n\to\mathbf{R}^n$ denote the projection on $\mathbf{R}^n$. Since

$$\pi_xD^+V(t_0,x_0)\subset D_x^+V(t_0,x_0)=:\{p_0\},$$

by (34) and (23) we conclude that

$$(p_t,p_x)\in D^*V(t_0,x_0)\Rightarrow p_x=p_0,\qquad p_t=H(t_0,x_0,-p_0).$$

Hence $D^+V(t_0,x_0)$ is a singleton. The conclusion follows from Proposition 3.10. □

COROLLARY 5.3. *Assume* (47) *that* g *is differentiable, and that the derivative* $V_x'(t_0,x_0)$ *does exist. Let* $\bar x$ *be an optimal solution of problem* (8).

Then for all $t\in[t_0,1]$, V *is differentiable at* $(t,\bar x(t))$ *and*

$$D^*V(t,\bar x(t))=\{V'(t,\bar x(t))\}.$$

Conversely, assume that $x:[t_0,1]\to\mathbf{R}^n$ *is a solution of* (6) *and that for every* $t\in[t_0,1]$, V *is differentiable at* $(t,x(t))$. *If the sets* $f(t,x,U)$ *are convex and compact, and*

$$-\left\langle\frac{\partial V}{\partial x}(t,x(t)),x'(t)\right\rangle=H\left(t,x(t),-\frac{\partial V}{\partial x}(t,x(t))\right)\quad \textit{a.e. in } [t_0,1],\tag{55}$$

then x *is optimal for problem* (8).

Proof. The first statement follows immediately from Proposition 5.2 and Theorem 4.5. To prove the second one, fix $\bar t\in[t_0,1]$ and let $\bar x:[\bar t,1]\to\mathbf{R}^n$ be an optimal solution of problem (8) with (t_0,x_0) replaced by $(\bar t,x(\bar t))$.

We already know that V is semiconcave. By Theorem 4.2, there exists $p(\bar t)\in\mathbf{R}^n$ such that

$$(H(\bar t,x(\bar t),p(\bar t)),-p(\bar t))=V'(\bar t,\bar x(\bar t)).$$

Since $\bar t\in[t_0,1]$ is arbitrary, assumption (55) and Theorem 4.1 end the proof. □

COROLLARY 5.4. *Under all assumptions of Theorem* 4.7 *assume* (47). *If problem* (8) *has a unique optimal solution* $\bar x$, *then* V *is differentiable at* $(t,\bar x(t))$ *for all* $t\in[t_0,1]$.

COROLLARY 5.5. *Under the hypotheses of Theorem* 4.6, *assume that* (47) *holds true and* g *is continuously differentiable. Then* $V(\cdot,\cdot)$ *is differentiable at* (t_0,x_0) *with the derivative* $(\partial V/\partial x)(t_0,x_0)$ *different from zero if and only if there is a unique optimal trajectory* $\bar x$ *for problem* (8) *satisfying* $g'(\bar x(1))\neq 0$.

Usually, the value function is not everywhere differentiable. However, that is always the case for "convex" problems, as we prove below (see also [4], [5], [9]).

PROPOSITION 5.6. *Assume that* (47) *holds true,* g *is convex, and for all* $t\in[0,1]$,

$$\text{Graph}\,(f(t,\cdot,U))\ \textit{is closed and convex.}\tag{56}$$

Then V is continuously differentiable on $[0,1]\times\mathbf{R}^n$ *and convex with respect to the second variable.*

Proof. By Theorem 2.1, assumption (56) yields that for every $(t_0, x_0)\in[0,1]\times\mathbf{R}^n$ there exists a solution $\bar{x}$ of the control system

$$x'=f(t,x(t),u(t)),\quad u(t)\in U,\quad x(t_0)=x_0,$$

satisfying $V(t_0,x_0)=g(\bar{x}(1))$.

Fix $t_0\in[0,1]$, $x_0, x_1\in\mathbf{R}^n$, $\lambda\in[0,1]$ and consider trajectories $x:[t_0,1]\to\mathbf{R}^n$ and $y:[t_0,1]\to\mathbf{R}^n$ such that $V(t_0,x_0)=g(x(1))$, $V(t_0,x_1)=g(y(1))$. Define the trajectory $z:[t_0,1]\to\mathbf{R}^n$ by $z(t)=\lambda x(t)+(1-\lambda)y(t)$. Then, using (56), we obtain that z is a solution of the control system (6) satisfying $z(t_0)=\lambda x_0+(1-\lambda)x_1$. Thus, by convexity of g,

$$V(t_0,\lambda x_0+(1-\lambda)x_1)\leqq g(z(1))\leqq\lambda V(t_0,x_0)+(1-\lambda)V(t_1,x_1)$$

and therefore $V(t_0,\cdot)$ is convex.

Next, as $V(t,\cdot)$ is both convex and semiconcave for all $t\in[0,1]$, Proposition 3.12 yields that $V(t,\cdot)$ is continuously differentiable on $\mathbf{R}^n$. The conclusion now follows from Proposition 5.2. □

6. Optimal feedback. One of the major issues of optimal control theory is to find an "equation" for optimal trajectories. Theorem 2.4 provides an inclusion formulation. However, in general, the set-valued map G is not regular enough to make us able to solve the inclusion (12). The situation is comparable to having an ordinary differential equation with a nonsmooth right-hand side: it may have solutions, but this solution cannot be obtained as, say, limits of Euler curves.

This is why we have to investigate regularity properties of G. In this section, we show that under the assumptions of Theorem 5.1, the feedback map G^{co} is upper semicontinuous and that so is G if we assume, in addition, that the sets $f(t,x,U)$ are closed.

The results of §§3 and 5 imply that under assumptions (47) the feedback maps $G:[0,1]\times\mathbf{R}^n\mapsto\mathbf{R}^n$ and $G^{\text{co}}:[0,1]\times\mathbf{R}^n\mapsto\mathbf{R}^n$ defined in § 2 are, respectively, equal to

$$G(t,x)=\{v\in f(t,x,U)\mid V_-^o(t,x)(1,v)=0\}$$

and

$$G^{\text{co}}(t,x)=\{v\in\overline{\text{co}}\,f(t,x,U)\mid V_-^o(t,x)(1,v)=0\}.$$

THEOREM 6.1. *Let us assume that* (47) *holds true. Then G^{co} has compact, nonempty images and is upper semicontinuous. The same holds true for the map G is we assume, in addition, that the sets $f(t,x,U)$ are closed.*

Proof. From Theorems 5.1 and 3.9 we know that for every $(t,x)\in[0,1[\times\mathbf{R}^n$ and every $\Theta\in\mathbf{R}^n$ the directional derivative $(\partial V/\partial(1,\Theta))(t,x)$ exists and is equal to the regularized lower derivative $V_-^o((t,x),(1,\Theta))$. Define the set-valued map

$$\hat{Q}:[0,1]\times\mathbf{R}^n\rightsquigarrow\mathbf{R}^n$$

by

$$\hat{Q}(t,x)=\{\Theta\in\mathbf{R}^n\mid V_-^o(t,x)(1,\Theta)\leqq 0\}.$$

From Proposition 3.5 we know that the set Graph $(\hat{Q})$ is closed. On the other hand, Proposition 2.3 implies that for every $v\in\overline{\text{co}}\,f(t,x,U)$, $(\partial V/\partial(1,v))(t,x)\geqq 0$. Thus

$$G(t,x)=\hat{Q}(t,x)\cap f(t,x,U),\qquad G^{\text{co}}(t,x)=\hat{Q}(t,x)\cap\overline{\text{co}}\,f(t,x,U).$$

This fact and the assumptions on f imply that the graphs of the set-valued maps G, G^{co} are closed. Furthermore, G^{co} takes its values in a compact set. From [2, p. 42] (see also [3, Chap. 1]) it follows that G and G^{co} are upper semicontinuous. □

COROLLARY 6.2. *Let us assume that* (47) *holds true and that the sets* $f(t, x, U)$ *are closed. If the map* G *is single-valued on a subset* $K \subset [0, 1] \times \mathbf{R}^n$, *then the function* $K \ni (t, x) \to G(t, x)$ *is continuous.*

A typical example of a nonlinear control system with closed convex images is the affine system

$$x' = f(x) + \sum_{i=1}^{k} u_i g_i(x), \qquad u_i \in [a, b],$$

where f and g_i are continuous functions from $\mathbf{R}^n$ to itself.

The feedback map G defined above, in general, does not have convex images because the map of directional derivatives is concave.

For this reason, in general, the feedback inclusion (12) is very difficult to investigate. When V happens to be differentiable and the sets $f(t, x, U)$ are closed and convex, then for obvious reasons the map G has convex compact images. Proposition 5.6 provides a sufficient condition for the continuous differentiability of V.

THEOREM 6.3. *Assume that* (47) *and* (56) *hold true and that* g *is convex. Then* G *has convex compact images and is upper semicontinuous. Furthermore, if for every* (t, x) *the set* $f(t, x, U)$ *is strictly convex, then* G *is single-valued and continuous.*

Proof. By Proposition 5.6, we know that V is continuously differentiable. This and the convexity of $f(t, x, U)$ yield that for all $(t, x) \in [0, 1[\times \mathbf{R}^n$,

$$G^{co}(t, x) = G(t, x) = f(t, x, U) \cap \{\Theta \in \mathbf{R}^n \mid V'(t, x)(1, \Theta) = 0\}$$

is convex. Theorem 6.1 ends the proof of the first statement. From Proposition 2.3 it follows that for all $(t, x) \in [0, 1[\times \mathbf{R}^n$,

$$v \in G(t, x) \Leftrightarrow v \in f(t, x, U) \text{ and}$$

$$\sup_{u \in U} \left\langle -\frac{\partial V}{\partial x}(t, x), f(t, x, u) \right\rangle = \left\langle -\frac{\partial V}{\partial x}(t, x), v \right\rangle.$$

This and strict convexity of $f(t, x, U)$ imply that G is single-valued. Corollary 6.2 completes the proof. □

7. Viability approach to optimal control. In this section, we provide an alternative approach to optimal trajectories based on viability techniques.

We first show the following characterization of optimal trajectories.

THEOREM 7.1. *Assume* (9). *Then a solution* $\bar{x}$ *of the control system* (6) *defined on the time interval* $[0, 1]$ *is optimal if and only if the function* $t \to (t, \bar{x}(t), V(0, \xi_0))$ *is a solution of the viability problem*

$$\begin{aligned} & t' = 1, \\ & x'(t) = f(t, x(t), u(t)), \qquad u(t) \in U \textit{ is measurable}, \\ (57) \qquad & z'(t) = 0, \\ & (t, x(t), z(t)) \in \text{Graph}(V) \quad \forall t \in [0, 1], \\ & t(0) = 0, \quad x(0) = \xi_0, \quad z(0) = V(0, \xi_0). \end{aligned}$$

Proof. We already observed that $\bar{x}(\cdot)$ is optimal if and only if the map $t \to V(t, \bar{x}(t))$ is constant on the time interval $[0, 1]$. On the other hand, $t \to (t, \bar{x}(t), z(t))$ is a solution of (57) if and only if $z(t)=V(t,\bar{x}(t))\equiv\text{const}$ and $\bar{x}(\cdot)$ is a solution of (6) satisfying $\bar{x}(0)=\xi_0$. □

Inclusion (57) is a viability problem which may be approached using many results of viability theory. Actually, the viability technique may be applied not only to the value function V but also to any continuous function W satisfying some inequalities from [17]. To state results in this direction, we need the following definition.

DEFINITION 7.2. Consider a continuous function $W:[0,1]\times\mathbf{R}^n\to\mathbf{R}$ and let $(t,x)\in[0,1]\times\mathbf{R}^n$. The contingent derivative of W at (t,x) in the direction $(w,v)\in\mathbf{R}\times\mathbf{R}^n$ is a subset of $\mathbf{R}$ defined by

$$\begin{aligned} &DW(t,x)(w,v)\\ (58)\qquad &:=\left\{u\in\mathbf{R}\;\middle|\;\liminf_{h\to0+,(w',v')\to(w,v)}\ \operatorname{dist}\left(u,\frac{W(t+hw',x+hv')-W(t,x)}{h}\right)=0\right\}.\end{aligned}$$

From [3, Chap. 6] it follows that the set $DW(t,x)(w,v)$ is closed and convex.

THEOREM 7.3. *Consider a continuous function* $W:[0,1]\times\mathbf{R}^n\to\mathbf{R}$ *and assume* (9). *If for every* $(t,x)\in[0,1]\times\mathbf{R}^n$,

$$0\in\{DW(t,x)(1,v)\,|\,v\in\overline{\text{co}}\,f(t,x,U)\},$$

then for all (t_0,x_0) *there exists a solution* $\bar{x}$ *of the differential inclusion*

$$x'(t)\in\overline{\text{co}}\,f(t,x(t),U)\quad\text{a.e. in }[t_0,1],\quad x(t_0)=x_0$$

such that $W(t,\bar{x}(t))\equiv W(1,\bar{x}(1))$.

Proof. It is not restrictive to assume that $t_0=0$. We extend W on $\mathbf{R}_+\times\mathbf{R}^n$ by setting for all $t>1$, $W(t,x)=W(1,x)$. Define the closed set $K=\text{graph}\,(W)$ and the map $F_1(t,x)=\{1\}\times\overline{\text{co}}\,f(t,x,U)\times\{0\}$. Set

$$\hat{F}(t,x)=\begin{cases}F_1(t,x) & \text{if } t\in[0,1[\\ \overline{\text{co}}\,(\{0\}\cup F_1(1,x)) & \text{if } t\geqq 1.\end{cases}$$

Then for every $(t,x)\in\mathbf{R}_+\times\mathbf{R}^n$, the contingent cone $T_K(t,x,W(t,x))$ to K at $(t,x,W(t,x))$ is equal to Graph $(DW(t,x))$ (see [3, Chap. 5]). Hence, by our assumption, for every $(t,x)\in[0,1[\times\mathbf{R}^n$ there exists $u\in F_1(t,x)$ such that $(1,u,0)\in T_K(t,x,W(t,x))$. Furthermore, for every $t\geqq1$ and $x\in\mathbf{R}^n$, we have $0\in\hat{F}(t,x)$. This proves that

$$\forall(t,x)\in\mathbf{R}_+\times\mathbf{R}^n,\quad \hat{F}(t,x)\cap T_K(t,x,W(t,x))\neq\emptyset.$$

By the assumptions, $\hat{F}$ is continuous and has closed convex images. Consequently, by the Haddad viability theorem[1] [20] (see also [1], [2]), the constrained system

$$\begin{aligned}&y'(t)\in\hat{F}(t,y(t))\quad\text{a.e.},\\ &y(t)\in K\quad\forall t\in[t_0,+\infty[,\\ &y(t_0)=(t_0,x_0,W(t_0,x_0))\end{aligned}$$

[1] This theorem states that, if an upper semicontinuous set-valued map F from a closed subset K of $\mathbf{R}^n$ into nonempty convex compact subsets of $\mathbf{R}^n$ has linear growth and satisfies the viability condition

$$\forall y\in K,\quad F(y)\cap T_K(y)\neq\emptyset,$$

then for every $y_0\in K$ there exists a solution of the differential inclusion

$$y'(t)\in F(y(t)),\quad y(0)=y_0,\quad y(t)\in K$$

defined on the whole half line $\mathbf{R}_+$.

has a solution $\bar{y}=(z_0, x, z):[t_0, +\infty[\to \mathbf{R}\times\mathbf{R}^n\times\mathbf{R}$. Then, from definition of K and $\hat{F}$, $z_0(t)=t$, $z(t)=W(t, x(t))$. On the other hand, $z'(t)=0$ almost everywhere in $[t_0, 1]$ and therefore $z\equiv\text{const}$. This ends the proof. □

THEOREM 7.4. *Consider a continuous function* $W:[0,1]\times\mathbf{R}^n\to\mathbf{R}$. *We assume that* f *does not depend on* t *and* (9) *holds true. If* $W(1,\cdot)=g(\cdot)$ *and*

$$\sup_{v\in\overline{\text{co}}\, f(x,U)} \inf D(-W)(t,x)(1,v)\leqq 0$$

(*where* $D(-W)(t,x)(1,v)\subset\mathbf{R}$ *is defined in* (58)), *then for every solution* $y(\cdot)=(t,x,z)(\cdot)$ *of*

$$(59)\qquad \begin{aligned} &t'=1,\\ &x'(t)=f(x(t),u(t)), \qquad u(t)\in U \textit{ is measurable},\\ &z'(t)=0,\\ &(t, x(t), z(t))\in \text{Graph}\,(W) \quad \forall t\in[0,1],\\ &t(0)=0, \quad x(0)=\xi_0, \quad z(0)=W(0,\xi_0), \end{aligned}$$

defined on the time interval $[0,1]$, *the trajectory* $x(\cdot)$ *is optimal for the problem* (7).

Proof. From [17] we deduce that W is nondecreasing along trajectories of (6). On the other hand, if $y(\cdot)=(t,x,z)(\cdot)$ is a solution of (59) defined on the time interval $[0,1]$, then $W(t,x(t))\equiv\text{const}$. □

8. Problem with endpoint constraints. In this section we investigate the case when the additional endpoint constraint is present:

$$x(1)\in K_1,$$

where K_1 is a given closed subset of $\mathbf{R}^n$. The corresponding value function is defined by

$$V(t_0, x_0)=\inf\{g(x(1))\,|\,x \text{ is a solution of (6) on } [t_0, 1], x(t_0)=x_0, x(1)\in K_1\}.$$

We observe that $V(t_0, x_0)=+\infty$ whenever no trajectory starting at x_0 at time t_0 hits K_1 at time one.

In this more general case, the value function may be discontinuous and we have either to develop a verification technique for a larger class of functions (some results in this direction were obtained in [17]) or to try to reduce the problem to a new one, where the data fits the Lipschitz framework. We will follow this second strategy and apply a penalization technique.

We provide only a convergence result showing that the problem with endpoint constraints may be approximated by free endpoint problems. Further developments are left to future work.

We impose assumptions (9) on the functions f and g, and we consider the family of penalized problems. With every $\varepsilon>0$, we associate the minimization problem

$$(\mathrm{P}_\varepsilon)\ \text{minimize}\left\{g(x(1))+\frac{1}{\varepsilon}\,\text{dist}\,(x(1), K_1)^2\,|\,x(\cdot) \text{ is a solution of (6)}, x(0)=\xi_0\right\}.$$

Define functions g_ε from $\mathbf{R}^n$ to $\mathbf{R}$ by

$$\forall x\in\mathbf{R}^n, \quad g_\varepsilon(x)=g(x)+\frac{1}{\varepsilon}\,\text{dist}\,(x, K_1)^2.$$

The value function V_ε corresponding to the problem (P_ε) is defined by (8) with g replaced by g_ε.

Since g_ε is locally Lipschitz, so does V_ε with the Lipschitz constant depending on ε. Hence the results obtained in the previous sections may be applied to V_ε.

Furthermore, if g is semiconcave, then, using Example 1 from § 3, we show that also the functions g_ε are semiconcave. This fact and Theorem 5.1 yield that, under assumptions (47), for every $\varepsilon > 0$ the value function V_ε is semiconcave on $[0, 1] \times \mathbf{R}^n$. Consequently, results concerning regularity of optimal feedback may be applied to penalized problems.

The aim of this section is to prove the convergence of V_ε to V.

THEOREM 8.1. *Assume* (9). *If the sets $f(t, x, U)$ are closed and convex, then for every $(t, x) \in [0, 1] \times \mathbf{R}^n$ the function $\mathbf{R}_+ \ni \varepsilon \to V_\varepsilon(t, x)$ is nonincreasing. Furthermore, for every $\varepsilon > 0$, $V_\varepsilon(t, x) \leqq V(t, x)$ and*

$$\lim_{\varepsilon \to 0+} V_\varepsilon(t, x) = V(t, x).$$

Proof. The first statement follows from the fact that the map $\varepsilon \to g_\varepsilon(x)$ is nonincreasing. The second one follows from the fact that $g(x(1)) = g_\varepsilon(x(1))$ whenever $x(1) \in K_1$.

Now fix $(t, x) \in [0, 1] \times \mathbf{R}^n$ and set $W(t, x) = \lim_{\varepsilon \to 0+} V_\varepsilon(t, x)$. Clearly, $W(t, x) \leqq V(t, x)$. To show the opposite, it is enough to consider the case $W(t, x) < +\infty$. Consider trajectory control pairs $(y^\varepsilon, u^\varepsilon)$ of control system (6) satisfying

$$V_\varepsilon(t, x) = g_\varepsilon(y^\varepsilon(1))$$

(they exist by Theorem 2.1). Then, by the relaxation and the parametrization theorems (see [2]), there exists a sequence $\varepsilon_n \to 0+$ and a trajectory $y(\cdot)$ of (6) defined on $[t, 1]$ such that $y^{\varepsilon_n} \to y$ uniformly on $[t, 1]$. On the other hand,

$$0 \leqq \operatorname{dist}(y^\varepsilon(1), K_1)^2 \leqq \varepsilon(W(t, x) - g(y^\varepsilon(1)))$$

and, therefore, taking the limit in the above inequality, we obtain $y(1) \in K_1$. Furthermore, from the inequality

$$V_\varepsilon(t, x) \geqq g(y^\varepsilon(1)),$$

we deduce that $W(t, x) \geqq g(y(1)) \geqq V(t, x)$. □

COROLLARY 8.2. *Under all assumptions of Theorem* 8.1, *consider a sequence $\varepsilon_n \to 0+$ and let $x^{\varepsilon_n}(\cdot)$ be an optimal solution to the problem* $(\mathrm{P}_{\varepsilon_n})$. *If problem* (P) *has at least one solution, then every cluster point $x(\cdot)$ of $\{x^{\varepsilon_n}(\cdot)\}$ in the metric of uniform convergence is an optimal solution of* (P).

Proof. Indeed, since (P) has a solution, by Theorem 8.1, for all $n > 0$,

$$g(x^{\varepsilon_n}(1)) \leqq g_{\varepsilon_n}(x^{\varepsilon_n}(1)) = V_{\varepsilon_n}(0, \xi_0) \leqq V(0, \xi_0) < +\infty$$

and, taking the limit, we deduce that $V(0, \xi_0) \geqq g(x(1)) \geqq V(0, \xi_0)$ and $x(1) \in K_1$. Thus x is optimal. □

REFERENCES

[1] J.-P. AUBIN, *Viability Theory*, Birkhäuser, Boston, Basel, Berlin, 1991.

[2] J.-P. AUBIN AND A. CELLINA, *Differential Inclusions*, Springer-Verlag, Berlin, 1984.

[3] J.-P. AUBIN AND H. FRANKOWSKA, *Set-Valued Analysis*, Birkhäuser, Boston, Basel, Berlin, 1990.

[4] V. BARBU AND G. DA PRATO, *Hamilton-Jacobi Equations in Hilbert Spaces*, Pitman, Boston, 1982.

[5] M. BARDI AND C. L. EVANS, *On Hopf's formula for solutions of Hamilton-Jacobi equations*, Nonlinear Anal., 8 (1984), pp. 1373-1381.

[6] L. BERKOVITZ, *Optimal feedback controls*, SIAM J. Control Optim., 27 (1989), pp. 991-1006.

[7] P. CANNARSA AND H. FRANKOWSKA, *Quelques charactérisations des trajectoires optimales en théorie de contrôle*, Note de CRAS, Série 1, Paris, 310 (1990), pp. 171-182.

[8] P. CANNARSA AND H. FRANKOWSKA, *Value function and optimality conditions for semilinear control problems*. Appl. Math. Optim., to appear.

[9] P. CANNARSA AND H. M. SONER, *On the singularities of the viscosity solutions to Hamilton-Jacobi-Bellman equations*, Indiana Univ. Math. J., 46 (1987), pp. 501-524.

[10] ———, *Generalized one-sided estimates for solutions of Hamilton-Jacobi equations and applications*, Nonlinear Anal., 13 (1989), pp. 305-323.

[11] F. H. CLARKE, *Optimization and Nonsmooth Analysis*, Wiley-Interscience, New York, (1983). Reprinted by CRM, Université de Montréal, 1989.

[12] F. H. CLARKE AND R. B. VINTER, *The relationship between the maximum principle and dynamic programming*, SIAM J. Control Optim. 25 (1987), pp. 1291-1311.

[13] M. G. CRANDALL AND P. L. LIONS, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., 277 (1983), pp. 1-42.

[14] 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.

[15] M. G. CRANDALL AND P. L. LIONS, *On existence and uniqueness of solutions of Hamilton-Jacobi equations*, Nonlinear Anal., 10 (1986), pp. 353-370.

[16] W. H. FLEMING AND W. H. RISHEL, *Deterministic and Stochastic Optimal Control*, Springer-Verlag, New York, 1975.

[17] H. FRANKOWSKA, *Optimal trajectories associated to a solution of contingent Hamilton-Jacobi equations*, Appl. Math. Optim., 19 (1989), pp. 291-311.

[18] ———, *Nonsmooth solutions of Hamilton-Jacobi-Bellman equation*, in Modeling and Control of Systems, A. Blaquiere, ed., Springer-Verlag, Berlin, 1988, pp. 131-147.

[19] ———, *Contingent cones to reachable sets of control systems*, SIAM J. Control Optim. 27 (1989), pp. 170-198.

[20] G. HADDAD, *Monotone trajectories of differential inclusions with memory*, Israel J. Math. 39 (1981), pp. 38-100.

[21] H. ISHII, *Uniqueness of unbounded viscosity solutions of Hamilton-Jacobi equations*, Indiana Univ. Math. J., 43 (1984), pp. 721-748.

[22] H. ISHII AND P. L. LIONS, *Viscosity solutions of fully nonlinear second order elliptic partial differential equation*, CEREMADE, 8820, 1988, preprint.

[23] P. L. LIONS, *Generalized Solutions of Hamilton-Jacobi equations*, Pitman, London, 1982.

[24] P. L. LIONS AND P. E. SOUGANIDIS, *Differential games, optimal control and directional derivatives of viscosity solutions of Bellman's and Isaaks' equations*, SIAM J. Control Optim., 23 (1985), pp. 566-583.

[25] N. N. SUBBOTINA, *The maximum principle and the superdifferential of the value function*, Problems Control Inform. Theory, 18 (1989), pp. 151-160.

[26] X. Y. ZHOU, *Maximum principle, dynamic programming and their connection in deterministic control*, J. Optim. Theory Appl., 65 (1990), pp. 363-373.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 6, pp. 1348-1372, November 1991

005

NEW SIZE×CURVATURE CONDITIONS FOR STRICT QUASICONVEXITY OF SETS*

GUY CHAVENT†

Abstract. Given a closed, not necessarily convex set D of a Hilbert space, the problem of the existence of a neighborhood $\mathcal{V}$ on which the projection on D is uniquely defined and Lipschitz continuous is considered, and such that the corresponding minimization problem has no local minima. After having equipped the set D with a family $\mathcal{P}$ of paths playing for D the role the segments play for a convex set, the notion of strict quasiconvexity of $(D, \mathcal{P})$ is defined, which will ensure the existence of such a neighborhood $\mathcal{V}$. Two constructive sufficient conditions for the strict-quasiconvexity of D are given, the R_G-size × curvature condition and the Θ-size × curvature condition, which both amount to checking for the strict positivity of quantities defined by simple formulas in terms of arc length, tangent vectors, and radii of curvature along all paths of $\mathcal{P}$. An application to the study of wellposedness and local minima of a nonlinear least squares problem is given.

Key words. projection theory, approximation theory, nonlinear least squares, inverse problems

AMS(MOS) subject classifications. 49A27, 52A50, 51K05

1. Introduction.

Let

$$F \text{ a Hilbert space,} \tag{1.1}$$

$$D \subset F \text{ a (not necessarily convex) subset} \tag{1.2}$$

be given. This paper is devoted to the study of the Hilbert projection on D. It is known (cf. [2]) that, when D is closed and bounded, the set of all $z \in F$ that admit a unique projection on D contains a dense countable intersection of open sets, called the Edelstein set, and defined by

$$\begin{aligned} ES = \{z \in F \mid \forall \varepsilon > 0, \exists \eta > 0 \text{ s.t. } X_j \in D, \|X_j - z\| \leqq d(z, D) + \eta, j = 0, 1 \\ \text{implies } \|X_0 - X_1\| \leqq \varepsilon\}. \end{aligned} \tag{1.3}$$

This result is not precise enough for the application to nonlinear least squares problems we have in mind, first because it gives only a generic result, and second because it does not give any insight into possible local minima on D of the "distance to z" function, whose presence or absence is critical when it comes to the actual numerical determination of the projection using an optimization algorithm.

Hence, it would be very useful to find conditions on D that ensure (i) that the Edelstein set contains some neighborhood $\mathcal{V}$ of D, and (ii) that ideally the $y \to d(y, z)$ function has no local minima on D whenever $z \in \mathcal{V}$. Necessary and sufficient conditions for point (i), involving the local curvature of D, have been extensively studied (cf., for example, [1] and [7]), but point (ii) has not received much attention.

The notion of the curvature of a set D is both delicate to define and difficult to apply to nonlinear least squares problems. That is why we will be looking only for *sufficient conditions for* (i) *and* (ii), which make it possible *to work only with the elementary notion of the curvature of a curve*, provided that the set D has been equipped with a collection $\mathcal{P}$ of paths P which will play for D the role of the segments for a convex set. This allows for an easy application to nonlinear least squares problems,

* Received by the editors May 7, 1990; accepted for publication (in revised form) November 6, 1990.

† Centre de la Recherche de Mathématique de la Décision, Université Paris Dauphine, 75775 Paris Cédex 16, France, and Institut National de Recherche en Informatique et en Automatique, BP 105, 78153 Le Chesnay Cédex, France.

where the paths $\mathcal{P}$ can be very naturally defined as the image of the segments of the admissible set by the mapping to be inverted. The notion of quasiconvex sets $(D, \mathcal{P})$ introduced in [6] following this line, together with a constructive sufficient condition (the γ-size × curvature condition) was a first answer to (i) and (ii): closed quasiconvex sets, which possess a neighborhood $\mathcal{V}$ on which the projection exists, is unique, and Lipschitz continuous, fully satisfy point (i), but not completely point (ii), as they do not eliminate local minima, which may exist provided their value is "large enough." Moreover, the γ-size × curvature condition introduced in [6] to recognize them does not seem to be very sharp, as, for example, it is far from recognizing all quasiconvex arcs of circles (cf. Figs. 3.2 and 3.3 below).

In this paper we remedy the above-mentioned weaknesses.

We first introduce the slightly stronger notion of a strictly quasiconvex set, which fully eliminates the possibility of local minima, as soon as the point to be projected is in the associated neighborhood $\mathcal{V}$.

Then we introduce the notion of a global radius of curvature R_G associated to the family of paths $\mathcal{P}$ and show that condition $R_G > 0$ (called the R_G-size × curvature condition) implies strict quasiconvexity of the set. This condition is shown to be sharp (see Figs. 3.2 and 3.3 below) as it recognizes exactly all strictly quasiconvex curves.

Third, we calculate a lower bound to R_G in terms of the usual radius of curvature R, the length Δ, and the deflection Θ associated to the family of paths $\mathcal{P}$; this, of course, leads to another sufficient condition for strict quasiconvexity of the set, called the Θ-size × curvature condition, which is also sharp because it recognizes exactly all strictly quasiconvex curves made up of one arc of a circle and one segment.

All geometrical quantities γ, R_G, R, Δ, and Θ are defined by infimum or supremum, over the collection of paths $\mathcal{P}$, of quantities easily computed on each path from the velocities and accelerations along the path.

Finally, we give an application to a nonlinear least squares problem in the case where the derivative of the nonlinear mapping to be inverted has a uniformly bounded pseudoinverse. A more detailed application to nonlinear least squares problems and their regularized versions can be found in [5].

2. Equipping the set with paths. The first step of our construction consists of choosing in the possibly nonconvex set D a collection $\mathcal{P}$ of paths P that will play for D the role the segments play for a convex set. We summarize here the corresponding definitions and notation, following paragraph 2.1 of [6].

DEFINITION 2.1 (Paths). A mapping $P:[0, L] \to D$ is a path if and only if

(2.1) $\nu \to P(\nu)$ is in $W^{2,\infty}([0, L])$;

(2.2) $\|P'(\nu)\|_F = 1$ for a.e. $\nu \in [0, L]$.

Note that this definition is slightly weaker than the one given in [6], where $\mathcal{C}^2$ regularity of the $\nu \to P(\nu)$ mapping was required. In fact, all results of [6] carry over when this new definition is used. We refer to [6] for a sufficient condition for a mapping to be reparametrizable in such a way that it is a path.

DEFINITION 2.2 (Attributes of a path). Let a path P be given. Then

(2.3) $\nu \in [0, L]$ is the arc-length along P;

(2.4) $\delta(P) \triangleq L$ is the length of P;

(2.5) $v(\nu) \triangleq P'(\nu)$ is the unit tangent vector to P at $P(\nu)$;

(2.6) $a(\nu) \triangleq P''(\nu)$ is the acceleration vector of P at $P(\nu)$;

(2.7) $\rho(\nu) \triangleq \|a(\nu)\|^{-1} \in IR^+ \cup \{+\infty\}$ is the radius of curvature of P at $P(\nu)$.

Because we will use only the parametrization by the arc length in this paper, we denote it simply by ν (note that in [6] the (reduced) arc length was denoted by $\bar{\nu}$, and all corresponding quantities wore bars, as quoted at the beginning of paragraph 3 of [6]; hence, $a(\nu)$ in this paper corresponds to $\bar{a}(\bar{\nu})$ in reference [6], etc.).

DEFINITION 2.3 (Collection of paths). A set of paths $\mathcal{P}$ is a *collection of paths* if and only if

(2.8) $\mathcal{P}$ *is made up of paths*;

(2.9) $\mathcal{P}$ *is complete*, i.e., $\forall X, Y \in D, X \neq Y, \exists P \in \mathcal{P}$, such that $P(0) = X$, $P(\delta(P)) = Y$;

(2.10) $\mathcal{P}$ is stable with respect to restriction, i.e., $\forall P \in \mathcal{P}, \forall \nu', \nu'' \in [0, \delta(P)], \nu' < \nu''$, the path $\tilde{P}: \nu \in [0, \nu'' - \nu'] \rightarrow P(\nu' + \nu)$ belongs to $\mathcal{P}$.

The set of paths $\mathcal{P}$ should contain the minimum number of paths that allow (2.9) to be satisfied (typically, one and only one path of $\mathcal{P}$ connects two given distinct points X and Y of D, but we do not put uniqueness in the requirements for a collection of paths because we do not exclude the possibility of choosing for $\mathcal{P}$ the minimum length paths between any two points of D, which may not be unique).

Note that hypothesis (2.1) implies that a path has a bounded curvature (by $\|a\|_\infty$), i.e., a strictly positive smallest radius of curvature. Hence we will be able to find paths $\mathcal{P}$ satisfying both Definitions 2.1 and 2.3 only if the set D itself is regular enough—intuitively if it also has a "bounded curvature."

DEFINITION 2.4 (Maximal paths). A subset $\mathcal{P}_M$ of $\mathcal{P}$ is said to be a collection of maximal paths for $\mathcal{P}$ if and only if

$$\mathcal{P} = \bigcup_{P \in \mathcal{P}_M} \{P' \mid P' \text{ is a subpath of } P\}. \tag{2.11}$$

Of course, such a $\mathcal{P}_M$ always exists because $\mathcal{P}_M = \mathcal{P}$ satisfies (2.11) by virtue of (2.10)! However, a more interesting case will be found when there exists a smallest collection of maximal paths: this will be the case in most of the applications to nonlinear least squares (see [6] and [5]), where $\mathcal{P}$ is the image, by a mapping φ, of the segments of a closed convex set C, so that $\mathcal{P}_M = \{\varphi([x, y]), x, y, \in \partial C\}$ obviously satisfies (2.11).

In the following, we will suppose that a collection of maximal paths $\mathcal{P}_M$ has been chosen.

3. Quasiconvex and strictly quasiconvex sets. Let the set D be equipped with a collection of paths $\mathcal{P}$. To a given point $z \in F$, a path $P \in \mathcal{P}$, and an arc length value $\nu \in [0, \delta(P)]$, we associate the number

$$k(z, P; \nu) = \langle z - P(\nu), a(\nu) \rangle_F \tag{3.1}$$

whose geometrical interpretation is given by (with the notation of Fig. 3.1)

$$k(z, P; \nu) = \frac{d(\nu)}{\rho(\nu)} \cos\theta = \frac{\overline{MH}}{\overline{MC}}. \tag{3.2}$$

We may then associate to z and P the number

$$k(z, P) = \operatorname*{sup\,ess}_{\nu \in [0, \delta(P)]} k(z, P; \nu) \tag{3.3}$$

and to z and $\eta > 0$ the number

$$k(z, \eta) = \sup_{P \in \mathcal{P}(z, \eta)} k(z, P), \tag{3.4}$$

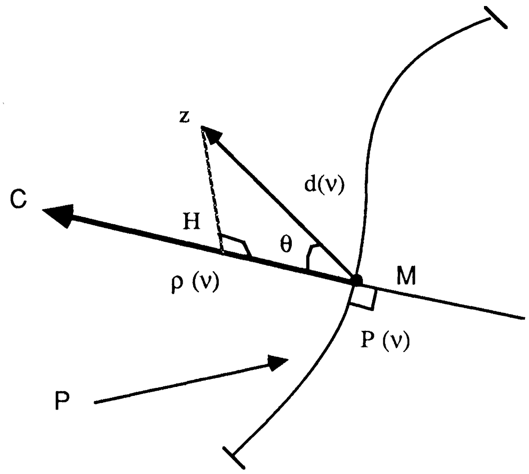


FIG. 3.1. *Notation for the geometrical interpretation of $k(z, P; \nu) = \langle z - P(\nu), a(\nu)\rangle_F$: C is the center of curvature of P at $P(\nu)$, and H is the orthogonal projection of z on the one-dimensional affine variety parallel to $a(\nu)$ and passing through $M = P(\nu)$.*

where the subcollection of paths $\mathcal{P}(z, \eta)$ is defined by

$$\mathcal{P}(z, \eta) = \{P \in \mathcal{P} \mid \|P(j) - z\|_F \leqq d(z, D) + \eta, j = 0, \delta(P)\}. \tag{3.5}$$

It will be convenient to define $k(z, 0)$ as the limit, when $\eta \to Ot$, of $k(z, 0)$.

We can now give the following definition.

DEFINITION 3.1 (Quasiconvexity of sets). A set $(D, \mathcal{P})$ is said to be *quasiconvex* if and only if

(i) $\mathcal{P}$ is a collection of paths;

(ii) There exists a neighborhood $\mathcal{V}$ of D in F, and a lower semicontinuous (l.s.c) function $\varepsilon : \mathcal{V} \to]0, +\infty]$ such that

$$\left.\begin{cases} z \in \mathcal{V} \\ 0 < \eta < \varepsilon(z) \end{cases}\right\} \Rightarrow k(z, \eta) < 1. \tag{3.6}$$

This is a slightly upgraded version of the definition of quasiconvexity given in [6] (where the $z \to \varepsilon(z)$ function was required to be continuous). All the results in [6] carry over when quasiconvexity is defined using Definition 3.1, and this upgraded definition allows us to prove the following proposition.

PROPOSITION 3.2. *Let $(D, \mathcal{P})$ be quasiconvex. Then there exists a largest open neighborhood $\mathcal{V}$ of D, and a largest l.s.c. function $\varepsilon : \mathcal{V} \to]0, +\infty[$ satisfying the definition of quasiconvexity.*

Proof. Let us denote by $\mathcal{V}_i$, ε_i, $i \in I$, all open neighborhoods and l.s.c. functions satisfying Definition 3.1. Then

$$\mathcal{V} = \bigcup_{i \in I} \mathcal{V}_i \tag{3.7}$$

is an open subset of F. If we denote by $\tilde{\varepsilon}_i$ the extension of ε_i to $\mathcal{V}$ by zero outside of $\mathcal{V}_i$, then $\tilde{\varepsilon}_i$ is l.s.c. as ε_i is l.s.c. on $\mathcal{V}_i$ and $\mathcal{V}_i$ is open. Define then

$$\varepsilon(z) = \sup_{i \in I} \tilde{\varepsilon}_i(z) \quad \forall z \in \forall, \tag{3.8}$$

which is l.s.c. as supremum of a family of l.s.c. functions. Hence $\mathcal{V}$, ε will satisfy the definition of quasiconvexity as soon as they satisfy (3.6), which we prove now. Let

$z \in \mathcal{V}$ and $0 < \eta < \varepsilon(z)$ be given, and set

$$\alpha = (\varepsilon(z) - \eta)/2 > 0.$$

From definition (3.8) of $\varepsilon(z)$, there exists $i_0 \in I$ such that

$$\tilde{\varepsilon}_{i0}(z) \geqq \varepsilon(z) - \alpha > \eta > 0.$$

This proves, as $\tilde{\varepsilon}_{i0}(z) > 0$, that $z \in \mathcal{V}_{i0}$. Hence $\mathcal{V}_{i0}$ and ε_{i0} satisfy the hypothesis of (3.6), so that

$$k(z, \eta) < 1,$$

which proves that $\mathcal{V}$ and $\varepsilon(z)$ satisfy (3.6). □

PROPOSITION 3.3. *Let $(D, \mathcal{P})$ be quasiconvex. Then the associated neighborhood $\mathcal{V}$ is included in the Edelstein set* (1.3), *and for any $z \in \mathcal{V}$, $0 < \eta < \varepsilon(z)$, and $P \in \mathcal{P}(z, \eta)$,*

(i) $\nu \to f(\nu) \triangleq \|P(\nu) - z\|_F^2$ *is strictly convex;*

(ii) $\nu \to d(\nu) \triangleq \|P(\nu) - z\|_F$ *is strictly quasiconvex.*

Proof. For $z \in \mathcal{V}$ and $P \in \mathcal{P}(z, \eta)$ we have

$$f''(\nu) \geqq 2(1 - k(z, \eta)) > 0 \quad \text{a.e.,}$$

which proves (i) and (ii), and also shows that the $\nu \to f(\nu) + \nu(\delta(P) - \nu)(1 - k(z, \eta))$ function is convex. Hence

$$f\left(\frac{\delta(P)}{2}\right) + \frac{\delta(P)^2}{4}(1 - k(z, \eta)) \leqq \frac{1}{2} f(0) + \frac{1}{2} f(\delta(P)).$$

However,

$$f\left(\frac{\delta(P)}{2}\right) \geqq d(z, D)^2, \qquad f(j) \leqq (d(z, D) + \eta)^2 \quad j = 0, \delta(P)$$

so that

$$\delta(P)^2 \leqq 4(1 - k(z, \eta))^{-1} \eta (2d(z, D) + \eta).$$

Let $X_j \in D$, $j = 0, 1$, be two η-projections of z on D:

$$\|X_j - z\| \leqq d(z, D) + \eta.$$

If we choose for P the path connecting X_1 and X_2, we see that $P \in \mathcal{P}(z, \eta)$; hence,

$$\|X_0 - X_1\|_F \leqq \delta(P) \leqq 2(1 - k(z, \eta))^{-1/2} \eta^{1/2} (2d(z, D) + \eta)^{1/2} \to 0 \text{ when } \eta \to 0,$$

which proves that z belongs to the Edelstein set. □

Proposition 3.3(ii) implies that the distance of any point z of $\mathcal{V}$ to any path of $\mathcal{P}$ having its extremities "not too far" from z is strictly quasiconvex.

Using the properties of the Edelstein set or the more precise results of [6], we can prove that, when $(D, \mathcal{P})$ is quasiconvex, the projection from $\mathcal{V}$ onto D is unique and Lipschitz continuous, and exists when D is closed. However, the "distance to z" function may still have, when $z \in \mathcal{V}$, parasitic local minima distinct from the global minimum. Of course, Proposition 3.3(ii) implies that these local minima are rejected at a distance of D larger than $d(z, D) + \varepsilon(z)$, but they may still exist and hence can be a source of trouble when we attempt to actually compute the projection using a gradient optimization algorithm.

We now introduce a new, stronger definition that will keep all the above-mentioned nice properties of quasiconvex sets with respect to projection, but will be much better behaved with respect to the problem of local minima.

DEFINITION 3.4 (Strict quasiconvexity of sets). A set $(D, \mathcal{P})$ is said to be *strictly quasiconvex* if and only if

(i) $(D, \mathcal{P})$ is quasiconvex for some neighborhood $\mathcal{V}$ and function $z \rightarrow \varepsilon(z)$;

(ii) $\mathcal{V}$ and ε satisfy, moreover,

$$\begin{array}{ll} z \in \mathcal{V} & \text{the "distance to } z\text{"} \\ P \in \mathcal{P} & \Rightarrow \text{function is strictly} \\ d(z, P) < d(z, D) + \varepsilon(z) & \text{quasiconvex along the path } P. \end{array} \tag{3.9}$$

The additional hypothesis (3.19) means (compare with Proposition 3.3(ii)) that the distance of any point z of $\mathcal{V}$ to any path of $\mathcal{P}$ that is "not too far" from z is strictly quasiconvex. We easily check that Proposition 3.2 extends to strictly quasiconvex sets.

PROPOSITION 3.5. *Let $(D, \mathcal{P})$ be strictly quasiconvex. Then there exists a largest open neighborhood $\mathcal{V}$ of D, and a largest l.s.c. function $\varepsilon : \mathcal{V} \rightarrow]0, +\infty]$ satisfying the definition of strict quasiconvexity.*

Of course, any neighborhood $\mathcal{V}$ associated by Definition 3.4 to a strictly quasiconvex set $(D, \mathcal{P})$ is included in the Edelstein set (1.3) by virtue of Proposition 3.3.

The following theorem summarizes the properties of strictly quasiconvex sets.

THEOREM 3.6 (Projection on strictly quasiconvex sets). *Suppose that*

$$(D, \mathcal{P}) \textit{ is strictly quasiconvex,} \tag{3.10}$$

and let $\mathcal{V}$, $\varepsilon(z)$ be the associated neighborhood and l.s.c. function.
Then

(i) *Uniqueness. For any $z \in \mathcal{V}$, there exists at most one projection $\hat{X}$ of z on D.*

(ii) *Local minima. If $z \in \mathcal{V}$ admits a (necessarily unique) projection $\hat{X}$ on D, the "distance to z" function has no parasitic local minimum on D distinct from $\hat{X}$.*

(iii) *Continuity. If $z_0, z_1 \in \mathcal{V}$ admit projections $\hat{X}_0, \hat{X}_1$ on D, and are close enough so that there exists $d \geqq 0$ satisfying*

$$\|z_0 - z_1\|_F + \max_{j=0,1} d(z_j, D) \leqq d < \min_{j=0,1} \{d(z_j, D) + \varepsilon(z_j)\}; \tag{3.11}$$

then, for any path P going from $\hat{X}_0$ to $\hat{X}_1$ we have

$$\|\hat{X}_0 - \hat{X}_1\| \leqq \delta(P) \leqq (1-k)^{-1} \|z_0 - z_1\|_F, \tag{3.12}$$

where $k < 1$ is defined by

$$\begin{aligned} &k = (k(z_0, \eta_0) + k(z_1, \eta_1))/2 \\ &0 < \eta_j = d - d(z_j, D) < \varepsilon(z_j), \qquad j = 0, 1. \end{aligned} \tag{3.13}$$

(iv) *Existence. If we suppose moreover that*

$$D \textit{ is closed in } F, \tag{3.14}$$

then any $z \in \mathcal{V}$ has a (unique) projection $\hat{X}$ on D, and any minimizing sequence X_n satisfies

$$\|X_n - \hat{X}\|_F \rightarrow 0 \tag{3.15}$$

and

$$\delta(P_n) \rightarrow 0 \textit{ where } P_n \textit{ is any path of } \mathcal{P} \textit{ going from } X_n \textit{ to } \hat{X}. \tag{3.16}$$

Proof. Properties (i), (iii), and (iv) result from the fact that strictly quasiconvex sets are quasiconvex sets, for which they have been proved to hold in [6]. Hence we are left with the proof of property (ii) on local minima. Let $X_0 \in D$ be one global minimum of the "distance to z" function, and suppose that this function admits a local minimum at $X_1 \neq X_0$. By definition of X_0 we have

$$\|z - X_0\| = d(z, D). \tag{3.17}$$

Then let $P \in \mathcal{P}$ connect X_0 to X_1. From (3.17) we obtain

$$d(z, P) = \|z - X_0\| = d(z, D) < d(z, D) + \varepsilon(z). \tag{3.18}$$

Then using property (3.9) of strictly quasiconvex sets we find that the "distance to z" function is strictly quasiconvex along the path P, which is impossible because this function has a global minimum at $\nu = 0$ and a local minimum at $\nu = \delta(P)$! □

We illustrate the notions of quasiconvex and strictly quasiconvex sets by considering the very simple case of a set D consisting in an arc of a circle of radius R and of length L. Simple geometric considerations show that

$$D \text{ is quasiconvex iff } \frac{L}{R} < 2\pi; \tag{3.19}$$

$$D \text{ is strictly quasiconvex iff } \frac{L}{R} < \pi. \tag{3.20}$$

We have illustrated in Fig. 3.2 for both cases the largest corresponding open neighborhood $\mathcal{V}$ whose existence is asserted by Propositions 3.2 and 3.5.

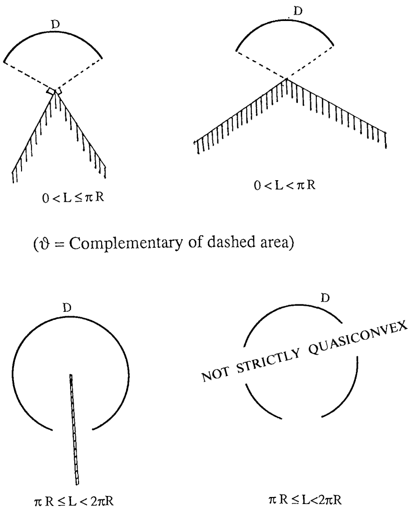


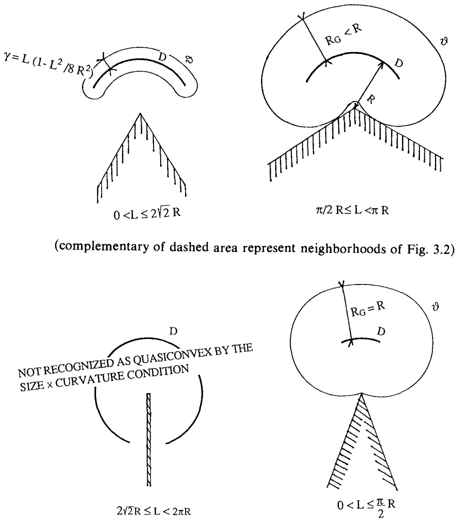
FIG. 3.2. *Largest open neighborhood $\mathcal{V}$ for quasiconvex ($L < 2\pi R$, left) and strictly quasiconvex ($L < \pi R$, right) arcs of circle.*

The first remark is that these neighborhoods are quite large; in particular, they are "infinite" on the "convex side" of the arc of the circle, and admit the center of the circle in their closure on the "concave side" of the arc of the circle.

However, if the quasiconvex neighborhood (Fig. 3.2, left) catches all points admitting a unique projection on D when $\pi R \leqq L < 2\pi R$, it misses many such points when $0 < L < \pi R$ (namely, all the points of the dashed area that are not on the axis of symmetry of the figure), whereas the strictly quasiconvex neighborhood (Fig. 3.2, right) does a much better job, as it catches in all cases exactly all points z admitting a unique projection on D with no local minima on D of the distance to z.

However, if it is simple to figure out the shape of these neighborhoods for sets D as simple as an arc of a circle, it becomes practically impossible for more complicated sets. Hence we need constructive sufficient conditions to recognize quasiconvexity and strict quasiconvexity, which will give us smaller cylindrical neighborhoods.

A first sufficient condition for quasiconvexity of sets was given in [6]: it associated, in a constructive way, a number γ to $(D, \mathscr{P})$, whose positiveness implied quasiconvexity of $(D, \mathscr{P})$ for a cylindrical neighborhood of D of size γ. This condition was not very precise: when applied to an arc of a circle, it read $L/R < 2\sqrt{2}$, and hence was far from recognizing all quasiconvex arcs of circles $(L/R < 2\pi)$, and the size γ of the corresponding cylindrical neighborhood was very small (see Fig. 3.3, left). We refer to this condition as the γ-size × curvature condition.



(complementary of dashed area represent neighborhoods of Fig. 3.2)

FIG. 3.3. *Cylindrical neighborhoods $\mathscr{V}$ given by the γ-size × curvature condition $(L < 2\sqrt{2}\, R$, left) and by the R_G-size × curvature condition $(L < \pi$, right).*

We turn now to the formulation of new sufficient conditions for strict quasiconvexity of sets, which will turn out to be much more precise that the γ-size × curvature condition: for arcs of circle they will give more (*strict* quasiconvexity) for less $(L/R < \pi$ instead of $L/R < 2\sqrt{2})$ together with a larger neighborhood! A preview of these results can be seen in Fig. 3.3, right.

4. The R_G-size × curvature condition: A new sufficient condition for strict quasiconvexity of sets.

We first introduce some new concepts associated with paths.

DEFINITION 4.1 (Affine normal subspace). Let a path P be given. Then, for any $\nu, \nu' \in [0, \delta(P)]$, $\nu \neq \nu'$, we define the *affine normal subspace* $N(\nu, \nu')$ to P at ν seen from ν' by

$$N(\nu, \nu') = \{z \in F \mid \langle z - P(\nu), \lambda v(\nu)\rangle_F \leqq 0 \quad \forall \lambda \in \mathbb{R}, \\ \nu + \lambda \in [\min(\nu, \nu'), \max(\nu, \nu')]\}. \tag{4.1}$$

This affine normal subspace is the dashed area passing through $P(\nu)$ on Fig. 4.1.

PROPOSITION 4.2. *Let a path P be given. Then, for any $\nu, \nu' \in [0, \delta(P)]$, $\nu \neq \nu'$, we have*

$$P(\nu) + a(\nu) \in N(\nu, \nu').$$

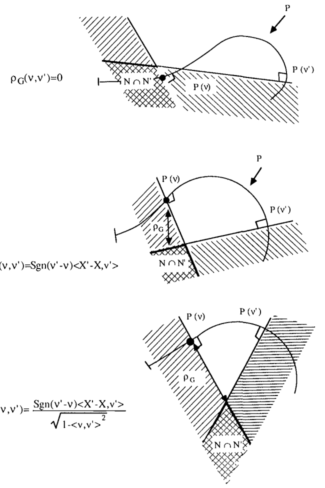

P
P(v')
ρG(v,v')=0
N∩N'
P(v)
P
P(v)
P(v')
ρG
ρG(v,v')=Sgn(v'-v)<X'-X,v'>
N∩N'
P(v)
P(v')
ρG
ρG(v,v')= Sgn(v'-v)<X'-X,v'> / √1-<v,v'>²
N∩N'

FIG. 4.1. *Examples of global radii of curvature for a path P.*

DEFINITION 4.3 (Global radius of curvature). Let a path P be given. Then, for any $\nu, \nu' \in [0, \delta(P)]$, $\nu \neq \nu'$, we define the global radius of curvature of P at ν seen from ν' by

$$\rho_G(\nu, \nu') = d(P(\nu), N(\nu, \nu') \cap N(\nu', \nu)) \in [0, +\infty] \tag{4.2}$$

with the natural convention that $\rho_G(\nu, \nu') = +\infty$ if $N(\nu, \nu') \cap N(\nu', \nu) = \emptyset$.

Note that $\rho_G(\nu, \nu') \neq \rho_G(\nu', \nu)$ in general. This global radius of curvature can be easily calculated.

PROPOSITION 4.4. *Let the path $P \in \mathcal{P}$ and $\nu, \nu' \in [0, \delta(P)]$, $\nu \neq \nu'$, be given, and denote*

$$\begin{aligned} X &= P(\nu), \qquad X' = P(\nu'), \\ v &= v(\nu), \qquad v' = v(\nu'), \\ N &= \operatorname{sgn}(\nu' - \nu)\langle X' - X, v'\rangle, \\ D &= \sqrt{1 - \langle v, v'\rangle^2}. \end{aligned} \tag{4.3}$$

Then $\rho_G(\nu, \nu')$ is given by

$$\rho_G(\nu, \nu') = \begin{cases} 0 & \text{if } N \leqq 0, \\ N & \text{if } N > 0 \quad \text{and} \quad \langle v, v'\rangle \leqq 0, \\ N/D & \text{if } N > 0 \quad \text{and} \quad \langle v, v'\rangle \geqq 0. \end{cases} \tag{4.4}$$

The proof of this formula is elementary, the basic ingredient being the projection of a point on a hyperplane. We have illustrated in Fig. 4.1 a few situations for the case of plane curves.

The relation with the usual (local) radius of curvature $\rho(\nu)$ is given by the following proposition.

PROPOSITION 4.5. *Let a path P be given. Then, for any $\nu \in [0, \delta(P)]$, there exists an open neighborhood $I(\nu)$ of ν in $[0, \delta(P)]$ such that, for any $\nu' \in I(\nu)$:*

$$\rho_G(\nu, \nu') = \frac{\operatorname{sgn}(\nu' - \nu)\langle X' - X, v'\rangle}{\sqrt{1 - \langle v, v'\rangle^2}}, \tag{4.5}$$

$$\rho_G(\nu', \nu) = \frac{\operatorname{sgn}(\nu' - \nu)\langle X' - X, v\rangle}{\sqrt{1 - \langle v, v'\rangle^2}} \tag{4.6}$$

and, for almost every $\nu \in [0, \delta(P)]$ *we have*

$$\rho_G(\nu, \nu') \to \rho(\nu), \tag{4.7}$$

$$\rho_G(\nu', \nu) \to \rho(\nu), \tag{4.8}$$

when $\nu' \to \nu$ *in* $I(\nu)$.

Proof. For any $v, \nu' \in [0, \delta(P)]$ we can always write

$$\operatorname{sgn}(\nu' - \nu)\langle X' - X, v'\rangle = |d\nu|\langle v_\theta, v'\rangle, \tag{4.9}$$

where

$$d\nu = \nu' - \nu, \qquad v_\theta = v(\nu + \theta d v) \quad 0 \leqq \theta \leqq 1.$$

Hence, given any $\nu \in [0, \delta(P)]$, we see that

$$\begin{matrix} \langle v, v'\rangle \to 1 \\ \langle v_\theta, v'\rangle \to 1 \end{matrix} \quad \text{when } d\nu \to 0, \tag{4.10}$$

which proves, in light of (4.4), that $\rho_G(\nu, \nu')$ is given by (4.5) when $d\nu$ is small enough. We could just as well have proved that $\rho_G(\nu', \nu)$ is given by (4.6) when $d\nu$ is small enough, which ends the proof of the existence of the interval $I(\nu)$ on which (4.5), (4.6) hold.

We now turn to the proof of (4.7). Let ν be a Lebesgue point for $a(\nu)$, $I(\nu)$ the corresponding interval, and $\nu' \in I(\nu)$, $\nu' \neq \nu$. Let us first transform the denominator of (4.5) then using the theorem of the median for v and v':

$$\langle v, v'\rangle = 1 - \frac{\delta^2}{2},$$

where

$$\delta = \|v' - v\|. \tag{4.11}$$

Hence the denominator of $\rho_G(\nu, \nu')$ may be rewritten as

$$\sqrt{1 - \langle v, v'\rangle^2} = \delta\left(1 - \frac{\delta^2}{4}\right)^{1/2} \tag{4.12}$$

and, using (4.9) for the numerator of ρ_G we obtain

$$\rho_G(\nu, \nu') = \frac{|d\nu|\langle v_\theta, v'\rangle}{\delta(1 - (\delta^2/4))^{1/2}}, \tag{4.13}$$

which, as $\langle v_\theta, v'\rangle \to 1$ and $\delta \to 0$, will prove (4.7) once we have found the limit of $\delta/|d\nu|$. But we have

$$\begin{aligned} \left|\frac{\delta}{|d\nu|} - \|a(\nu)\|\right| &= \left|\frac{1}{|d\nu|}\left\|\int_0^{d\nu} a(\nu + \varepsilon)\, d\varepsilon\right\| - \|a(\nu)\|\right| \\ &\leqq \frac{1}{|d\nu|}\left\|\int_0^{d\nu} [a(\nu + \varepsilon) - a(\nu)]\, d\varepsilon\right\| \\ &\leqq \frac{1}{|d\nu|}\int_0^{d\nu} \|a(\nu + \varepsilon) - a(\nu)\|\, d\varepsilon, \end{aligned}$$

which, by definition of the Lebesgue points, tends to zero when $d\nu \to 0$ (see, for example, Theorem 8.8 of [8]). Hence

$$\frac{\delta}{|d\nu|} \to \|a(\nu)\|, \quad \text{when } d\nu \to 0. \tag{4.14}$$

This shows that (4.7) holds at all Lebesgue points of $[0, \delta(P)]$. Because almost every point of $[0, \delta(P)]$ is a Lebesgue point, we have proved (4.7).

Then (4.8) follows immediately because

$$\rho_G(\nu, \nu') - \beta(d\nu) \leqq \rho_G(\nu', \nu) \leqq \rho_G(\nu, \nu') + \beta(d\nu), \tag{4.15}$$

with

$$\beta(d\nu) = \frac{|\langle X' - X, v - v' \rangle|}{\sqrt{1 - \langle v, v' \rangle^2}} = \frac{|d\nu| |\langle v_{\theta'}, v - v' \rangle|}{\sqrt{1 - \langle v, v' \rangle^2}},$$

where $0 < \theta' < 1$. Using (4.11) and (4.12) we obtain

$$\beta(d\nu) \leqq \frac{|d\nu|}{(1 + (\delta^2/4))^{1/2}} \to 0, \quad \text{when } d\nu \to 0. \tag{4.16}$$

This ends the proof of Proposition 4.5.

We are now naturally led to the following definition.

Definition 4.6 (Smallest global radii of curvature of a path). Let a path P be given. Then we define

$$\begin{aligned} R(P) &= \inf_{\nu \in [0, \delta(P)]} \rho(\nu) && \text{(smallest radius of curvature of } P\text{)}, \\ R_G(P) &= \inf_{\nu, \nu' \in [0, \delta(P)]} \rho_G(\nu, \nu') && \text{(smallest global radius of curvature of } P\text{)}. \end{aligned} \tag{4.17}$$

Of course, $R_G(P)$ is easily related to $R(P)$ using Proposition 4.5.

Proposition 4.7. *Let a path P be given. Then*

$$0 \leqq R_G(P) \leqq R(P) \leqq +\infty. \tag{4.18}$$

Remark 4.8. Unlike the number $\gamma(P)$ used in the size × curvature condition developed in [4] and [6] and recalled at the end of § 3, the number $R_G(P)$ depends only on the shape of the path P, and in no way on its parametrization (that is why in this paper we use only the parametrization by arc length, which is the most convenient for the formulas).

The significance of the smallest global radius of curvature comes from the following proposition.

Proposition 4.9. *Let a path $P \in \mathcal{P}$ and $z \in F$ be given. If*

$$R_G(P) > 0, \qquad d(z, P) < R_G(P),$$

then the $\nu \to d(\nu) = \|P(\nu) - z\|_F$ function is strictly quasiconvex.

Proof. Let $f(\nu) = d(\nu)^2$, and $\nu_0 \in [0, \delta(P)]$ be the value for which d and f attain their minimum on $[0, \delta(P)]$:

$$f(\nu_0) \leqq f(\nu), \quad \forall \nu \in [0, \delta(P)],$$

which implies

$$f'(\nu_0)\lambda \geqq 0, \quad \forall \lambda \in \mathbb{R}, \qquad \nu_0 + \lambda \in [0, \delta(P)];$$

i.e.,

$$\langle z-P(\nu_0), \lambda v(\nu_0)\rangle \leqq 0, \quad \forall \lambda \in \mathbb{R}, \qquad \nu_0+\lambda \in [0, \delta(P)],$$

which shows that

$$z \in N(\nu_0, \nu') \quad \text{for any } \nu' \in [0, \delta(P)], \nu' \neq \nu_0.$$

Suppose now that $\nu \to d(\nu)$ is not strictly quasiconvex. Then necessarily $d(\nu)$ has at least one local maximum for some value ν_1 in the *open interval* $]0, \delta(P)[$. Of course, $\nu \to f(\nu)$ also has a local maximum at the same $\nu_1 \in]0, \delta(P)[$, which implies that $f'(\nu_1)=0$; i.e.,

$$\langle z-P(\nu_1), v(\nu_1)\rangle = 0.$$

However, the preceding equation may be rewritten as

$$z \in N(\nu_1, \nu') \quad \text{for any } \nu' \in [0, \delta(P)], \nu' \neq \nu_1.$$

Hence we see that $z \in N(\nu_0, \nu_1) \cap N(\nu_1, \nu_0)$, which shows that

$$\rho_G(\nu_0, \nu_1) \leqq \|P(\nu_0)-z\|_F = d(z, P),$$

and, using definition (4.17):

$$R_G(P) \leqq \rho_G(\nu_0, \nu_1) \leqq d(z, P),$$

which contradicts the hypothesis $d(z, P) < R_G(P)$. This proves the strict quasiconvexity of d. □

DEFINITION 4.10 (Smallest radii of curvature of a set). We associate to a set $(D, \mathcal{P})$

$$\text{(4.19)} \qquad \begin{aligned} R(D) &= \inf_{P \in \mathcal{P}_M} R(P) && \text{(smallest radius of curvature in } D), \\ R_G(D) &= \inf_{P \in \mathcal{P}_M} R_G(P) && \text{(smallest global radius of curvature in } D), \end{aligned}$$

which, using (4.18), satisfy

$$\text{(4.20)} \qquad 0 \leqq R_G(D) \leqq R(D) \leqq +\infty.$$

This allows us to define a new "size × curvature" condition for the set D.

DEFINITION 4.11 (R_G-size × curvature condition). The set $(D, \mathcal{P})$ is said to satisfy a R_G-size × curvature condition if and only if

$$\text{(4.21)} \qquad R_G(D) > 0.$$

We give now our first main result.

THEOREM 4.12. *Let $(D, \mathcal{P})$ be given. If*

$$\text{(4.22)} \qquad R_G(D) > 0 \quad (R_G\textit{-size} \times \textit{curvature condition}),$$

then D is strictly quasiconvex, with a cylindrical neighborhood $\mathcal{V}$ given by

$$\text{(4.23)} \qquad \mathcal{V} = \{z \in F \mid d(z, D) < R_G(D)\},$$

and an $\varepsilon(z)$ function defined, for any $z \in \mathcal{V}$, by

$$\text{(4.24)} \qquad \varepsilon(z) = R_G(D) - d(z, D) > 0.$$

Hence Theorem 3.6 (on the properties of the projection) applies, the Lipschitz constant being given by (compare with (3.11)-(3.13))

$$\text{(4.25)} \qquad \|\hat{X}_0 - \hat{X}_1\|_F \leqq \delta(P) \leqq \left(1 - \frac{d}{R(D)}\right)^{-1} \|z_0 - z_1\|_F$$

as soon as z_0, z_1 are close enough so that there exists d satisfying

$$\|z_0 - z_1\|_F + \max_{j=0,1} d(z_j, D) \leqq d < R_G(D). \tag{4.26}$$

Proof. We check first that $\mathcal{V}$ and ε defined by (4.23) and (4.24) satisfy Definition 3.1 of quasiconvex sets. Let z, η, and P be given such that

$$d(z, D) < R_G(D) \quad (\text{i.e., } z \in \mathcal{V}), \tag{4.27}$$

$$0 < \eta < \varepsilon(z) = R_G(D) - d(z, D), \tag{4.28}$$

$$\|P(j) - z\| \leqq d(z, D) + \eta \qquad j = 0, \delta(P). \tag{4.29}$$

Hence $\|P(j) - z\| < R_G(D)$, $j = 0, \delta(P)$ which implies that $d(z, P) < R_G(D) \leqq R_G(P)$. Then by Proposition 4.9 we see that the $\nu \to d(\nu) = \|P(\nu) - z\|_F$ is strictly quasiconvex, and hence, using (4.29),

$$d(\nu) \leqq \max_{j=0,\delta(P)} \|P(j) - z\| \leqq d(z, D) + \eta, \quad \forall \nu \in [0, \delta(P)].$$

However, from definitions (4.6) and (4.10) we see that

$$\rho(\nu) \geqq R(P) \geqq R(D) \quad \text{for a.e. } \nu \text{ in } [0, \delta(P)].$$

Plugging the two last bounds into (3.2) yields

$$k(z, P; \nu) \leqq \frac{d(z, D) + \eta}{R(D)} \quad \text{for a.e. } \nu \text{ in } [0, \delta(P)],$$

which, as this bound is independent of the path P provided it belongs to $\mathcal{P}(z, \eta)$, proves that (cf. (3.4))

$$k(z, \eta) \leqq \frac{d(z, D) + \eta}{R(D)} < 1, \tag{4.30}$$

which shows that D is quasiconvex. The strict quasiconvexity of D results then directly from Proposition 4.9, and the formulas (4.25) and (4.26) follow immediately from (3.11)-(3.12) and (4.30). □

To illustrate this new sufficient condition, we check how well it recognizes strictly quasiconvex arcs of circles of radius R and length L. A simple calculation yields

$$R_G(D) = \begin{cases} R & 0 < \dfrac{L}{R} < \dfrac{\pi}{2}, \\ R \sin \dfrac{L}{R} & \dfrac{\pi}{2} \leqq \dfrac{L}{R} \leqq \pi, \\ 0 & \pi \leqq \dfrac{L}{R} < 2\pi, \end{cases} \tag{4.31}$$

which is strictly positive as long as

$$0 < \frac{L}{R} < \pi, \tag{4.32}$$

(compare with (3.20)).

Hence the *R_G-size × curvature condition recognizes exactly all strictly quasiconvex arcs of circles*! We have illustrated in Fig. 3.3, right, the neighborhoods $\mathcal{V}$ generated by (4.31), which turn out to be the *largest cylindrical neighborhoods* included in the *largest strictly quasiconvex neighborhood* (depicted as the complement of the dashed area). As mentioned at the end of § 3, comparison with the left part of Fig. 3.3 shows that the R_G-size×curvature condition is much more precise than the old γ-size× curvature condition.

In the general case, the R_G-size×curvature condition cannot be expected to be optimal (i.e., necessary for strict quasiconvexity), as nothing prevents equipping a convex set with paths $\mathcal{P}$ having small radii of curvature! However, it is reasonable to conjecture that the R_G-size×curvature condition becomes necessary in the case where the collection of paths $\mathcal{P}$ is made of minimum length paths of D. This matter will be discussed in a forthcoming paper.

5. Obtaining lower bounds on the global radius of curvature. Given a path $P \in \mathcal{P}$, we will try in this section to obtain lower bounds on its global radius of curvature $R_G(P)$.

In particular, we would like to substantiate the intuitive feeling that "arcs of circles are the worst paths," i.e., that for any path P we should have $R_G(P) = R(P)$ as soon as $\delta(P) \leqq (\pi/2)R(P)$ and $R_G(P) > 0$ as soon as $\delta(P) < \pi R(P)$, as suggested by Fig. 3.3 (right) for the case of an arc of a circle. To do that, we need to recall one (very natural) attribute of the paths.

DEFINITION 5.1 (Deflection). Given a path $P \in \mathcal{P}$, and $\nu, \nu' \in [0, \delta(P)]$, the *deflection of P between ν and ν' is*

$$\theta(\nu, \nu') = \operatorname{Arg}\cos\langle v(\nu), v(\nu')\rangle, \tag{5.1}$$

and the *largest deflection of P* is

$$\Theta(P) = \max \theta(\nu, \nu') \qquad \nu, \nu' \in [0, \delta(P)]. \tag{5.2}$$

Of course, $\theta(\nu, \nu')$ satisfies

$$\begin{aligned} &\theta(\nu, \nu') \in [0, \pi] && \forall \nu, \nu' \in [0, \delta(P)],\\ &\theta(\nu, \nu) = 0 && \forall \nu \in [0, \delta(P)],\\ &\theta(\nu, \nu') = \theta(\nu', \nu) && \forall \nu, \nu' \in [0, \delta(P)]. \end{aligned} \tag{5.3}$$

Given $\nu, \nu' \in [0, \delta(P)]$ and any function $g: [\nu, \nu'] \to \mathbb{R}$ we denote by $\operatorname{var} g$ the total variation (when it exists!) of g over the $[\nu, \nu']$ interval defined by

$$\operatorname{var} g = \sup_{\substack{N \in \mathbb{N} \\ \min(\nu,\nu') \leqq t_0 \leqq t_1 \cdots \leqq t_N = \max(\nu,\nu')}} \sum_{i=1}^{N} |g(t_i) - g(t_{i-1})|. \tag{5.4}$$

The regularity of the deflection function and its relation with the radius of curvature are given by the following lemma.

LEMMA 5.2. *Let $P \in \mathcal{P}$ be given. Then for any $\nu' \in [0, \delta(P)]$, the $\nu \to \theta(\nu, \nu')$ deflection function is absolutely continuous and has a bounded variation over* $[0, \delta(P)]$. *Hence $\partial\theta/\partial\nu(\nu, \nu')$ exists almost everywhere on* $[0, \delta(P)]$, $\partial\theta/\partial\nu(\cdot, \nu') \in L^1([0, \delta(P)])$, *and the usual formulas hold*

$$\theta(\nu, \nu') = \int_{\nu'}^{\nu} \frac{\partial\theta}{\partial\nu}(t, \nu')\, dt, \tag{5.5}$$

$$\operatorname{var} \theta(\cdot, \nu') = \int_{\nu}^{\nu'} \left|\frac{\partial\theta}{\partial\nu}(t, \nu')\right| dt. \tag{5.6}$$

Moreover,

$$(5.7)\qquad \left|\frac{\partial\theta}{\partial\nu}(\nu,\nu')\right| \leqq \|a(\nu)\| = \frac{1}{\rho(\nu)} \quad \text{for a.e. } \nu\in[0,\delta(P)],$$

which implies that in fact $\partial\theta/\partial\nu(\cdot,\nu')\in L^\infty([0,\delta(P)])$.

Proof. Let $P\in\mathcal{P}$ and $\nu'\in[0,\delta(P)]$ be given.

Step 1. There exists $\Delta\nu>0$ such that, for any $\nu_1,\nu_2\in[0,\delta(P)]$ and satisfying $|\nu_1-\nu_2|\leqq\Delta\nu$, we have

$$(5.8)\qquad \theta(\nu_1,\nu_2)\leqq\pi/3$$

and

$$(5.9)\qquad |\theta(\nu_1,\nu')-\theta(\nu_2,\nu')|\leqq\frac{\|v_1-v_2\|}{\sqrt{\cos\theta(\nu_1,\nu_2)}}\leqq\sqrt{2}\,\|v_1-v_2\|.$$

The existence of $\Delta\nu>0$ such that (5.8) holds results immediately from the uniform continuity of $\theta(\cdot,\cdot)$ over $[0,\delta(P)]\times[0,\delta(P)]$. To prove (5.9) we first use the triangular inequality for the curvilinear triangle on the unit sphere having $v(\nu_1)$, $v(\nu_2)$ and $v(\nu')$ as vertices:

$$(5.10)\qquad |\theta(\nu_1,\nu')-\theta(\nu_2,\nu')|\leqq\theta(\nu_1,\nu_2).$$

Using the theorem of the median we obtain, as in Proposition 4.5,

$$\cos\theta(\nu_1,\nu_2)=\langle v_1,v_2\rangle=1-\frac{\|v_1-v_2\|^2}{2}.$$

On the other hand, a Taylor-MacLaurin development of $\cos\theta$ to the order 2 yields

$$\cos\theta(\nu_1,\nu_2)=1-\frac{\theta(\nu_1,\nu_2)^2}{2}\cos(\alpha\theta(\nu_1,\nu_2))\qquad 0\leqq\alpha\leqq1.$$

However, from (5.8) we see that

$$\cos(\alpha\theta(\nu_1,\nu_2)\geqq\cos\theta(\nu_1,\nu_2)\geqq\tfrac{1}{2},$$

which finishes the proof of (5.9).

Step 2. We prove that $\theta(\cdot,\nu')$ is absolutely continuous and has a bounded variation, which automatically implies the existence of $\partial\theta/\partial\nu$ almost everywhere, and hence proves (5.5) and (5.6) (see, for example, Theorems 8.14 and 8.18 of [8]).

We begin by proving the absolute continuity. Let $\varepsilon>0$ be given, and let (α_i,β_i) $i=1,2,\cdots,N$ be disjoint segments of the interval $[0,\delta(P)]$ satisfying

$$\beta_i-\alpha_i\leqq\Delta\nu\quad\forall i=1,\cdots,N.$$

Then we see from (5.9) that

$$\sum_{i=1}^N|\theta(\beta_i,\nu')-\theta(\alpha_i,\nu')|\leqq\sqrt{2}\sum_{i=1}^N\|v(\beta_i)-v(\alpha_i)\|.$$

However, since $P\in W^{2,\infty}([0,\delta(P)]$ we have

$$\|v(\beta_i)-v(\alpha_i)\|\leqq\int_{\alpha_i}^{\beta_i}\|a(t)\|\,dt\leqq\|a\|_\infty(\beta_i-\alpha_i),$$

so that

$$\sum_{i=1}^N|\theta(\beta_i,\nu')-\theta(\alpha_i,\nu')|\leqq\sqrt{2}\,\|a\|_\infty\sum_{i=1}^N(\beta_i-\alpha_i)\leqq\varepsilon$$

as soon as the α_i, β_i satisfy

$$\sum_{i=1}^{N} (\beta_i - \alpha_i) \leqq \min\left\{\Delta\nu, \frac{\varepsilon}{\sqrt{2}\,\|a\|_\infty}\right\}.$$

We now prove that $\theta(\cdot, \nu')$ has a bounded variation over $[0, \delta(P)]$. Let $t_i\, i = 0, 1, \cdots, N$ be given such that

$$0 \leqq t_0 < t_1 < \cdots < t_N \leqq \delta(P).$$

We can always *add a finite number of points* to obtain a new subdivision

$$0 \leqq t'_0 < t'_1 < \cdots < t'_{N'} \leqq \delta(P)$$

(with $N' \geqq N$!) such that

$$|t'_i - t'_{i-1}| \leqq \Delta\nu, \qquad i = 1, 2, \cdots, N'.$$

Then, of course, the triangular inequality implies

$$\sum_{i=1}^{N} \|\theta(t_i, \nu') - \theta(t_{i-1}, \nu')| \leqq \sum_{i=1}^{N'} \|\theta(t'_i, \nu') - \theta(t'_{i-1}, \nu')|.$$

However, from (5.9) we find that for all $i = 1, \cdots, N'$ we have

$$|\theta(t'_i, \nu') - \theta(t'_{i-1}, \nu')| \leqq \sqrt{2}\, \|a\|_\infty (t'_1 - t'_{i-1}),$$

which shows that

$$\sum_{i=1}^{N} |\theta(t_i, \nu') - \theta(t_{i-1}, \nu')| \leqq \sqrt{2}\, \|a\|_\infty \delta(P)$$

independently of the positions of the points t_i, $i = 0, \cdots, N$, and of their number N.

Step 3. We prove (5.7). Let $\nu \in [0, \delta(P)]$ be a *Lebesgue point* for both $a(\cdot) \in L^\infty[0, \delta(P)]$ and $\partial\theta/\partial\nu(\cdot, \nu') \in L^1[0, \delta(P)]$ (almost every point of $[0, \delta(P)]$ has this property!). Hence we know (see (4.14) in the proof of Proposition 4.5) that

$$\|a(\nu)\| = \lim_{d\nu \to 0} \frac{\|v(\nu + d\nu) - v(\nu)\|}{|d\nu|}$$

and

$$\frac{\partial\theta}{\partial\nu}(\nu, \nu') = \lim_{d\nu \to 0} \frac{1}{d\nu} \int_{\nu}^{\nu + d\nu} \frac{\partial\theta}{\partial\nu}(t, \nu')\, dt,$$

which, using (5.5), may be rewritten as

$$\frac{\partial\theta}{\partial\nu}(\nu, \nu') = \lim_{d\nu \to 0} \frac{\theta(\nu + d\nu, \nu') - \theta(\nu, \nu')}{d\nu}.$$

If we choose $d\nu \leqq \Delta\nu$ of Step 1, we see from (5.9) that

$$\left|\frac{\theta(\nu + d\nu, \nu') - \theta(\nu, \nu')}{d\nu}\right| \leqq \frac{\|v(\nu + d\nu) - v(\nu)\|}{|d\nu|} \frac{1}{\sqrt{\cos\theta(\nu + d\nu, \nu)}},$$

which proves (5.7) when $d\nu \to 0$ as $\theta(\nu + d\nu, \nu) \to 0$. This ends the proof of Lemma 5.2. □

So much for the properties of the deflection function. We come now to our purpose, namely, obtaining a lower bound on $R_G(P)$, which reduces to finding a lower bound on $\rho_G(\nu, \nu')$ independent of ν and ν'. Looking at (4.4) giving $\rho_G(\nu, \nu')$, we see that the two pieces entering in this formula are

$$\text{sgn}\,(\nu'-\nu)\langle X'-X, v'\rangle \quad \text{and} \quad \sqrt{1-\langle v, v'\rangle 2}, \tag{5.11}$$

which are related to the deflection θ by

$$\text{sgn}\,(\nu'-\nu)\langle X'-X, v'\rangle = \int_{\min(\nu,\nu')}^{\max(\nu,\nu')} \cos\theta(t, \nu')\,dt \tag{5.12}$$

and

$$\sqrt{1-\langle v, v'\rangle 2} = \sin\theta(\nu, \nu'). \tag{5.13}$$

We now concentrate on obtaining a lower bound for $\text{sgn}\,(\nu', \nu)\langle X'-X, v'\rangle$. Our basic tool for that is the following lemma.

LEMMA 5.3. *Let $\nu, \nu' \in [0, \delta(P)]$ be given. Then the following inequalities hold*:

$$\text{var}\,\theta(\cdot, \nu') \leqq \int_{\min(\nu,\nu')}^{\max(\nu,\nu')} \frac{d\nu}{\rho(\nu)}, \tag{5.14}$$

$$\text{sgn}\,(\nu'-\nu)\langle X'-X, v'\rangle \geqq |\nu'-\nu| \cos\bar\theta + \bar R\,\text{var}\,\{\sin\theta(\cdot, \nu') - \theta(\cdot, \nu')\cos\bar\theta\}, \tag{5.15}$$

where

$$\bar R = \inf\text{ess}\,\rho(t) \qquad t\in[\nu, \nu'], \tag{5.16}$$

$$\bar\theta = \sup\theta(t, \nu') \qquad t\in[\nu, \nu']. \tag{5.17}$$

Proof. Let us consider the case where $\nu \leqq \nu'$ (the proof is similar if $\nu \geqq \nu'$). We remark first that (5.14) follows immediately from (5.6) and (5.7). Then from (5.12) and (5.17), we find that

$$\langle X'-X, v'\rangle = \int_\nu^{\nu'} (\cos\theta(t, v') - \cos\bar\theta)\,dt + (\nu'-\nu)\cos\bar\theta. \tag{5.18}$$

From Lemma 5.2 we know that

$$\rho(t)\left|\frac{d\theta}{d\nu}(t, \nu')\right| \leqq 1 \quad \text{for almost every } t \text{ of } [\nu, \nu']. \tag{5.19}$$

Plugging then (5.19) into (5.18) yields, as $\cos\theta(t, \nu') - \cos\bar\theta \geqq 0$ for all $t\in[\nu, \nu']$ and $\rho(t) \geqq \bar R$ for all $t\in[\nu, \nu']$:

$$\begin{aligned}\langle X'-X, v'\rangle &\geqq \bar R \int_\nu^{\nu'} (\cos\theta(t, \nu') - \cos\bar\theta)\left|\frac{\partial\theta}{\partial\nu}(t, \nu')\right| dt + (\nu'-\nu)\cos\bar\theta \\ &= \bar R \int_\nu^{\nu'} \left|\frac{\partial}{\partial\nu}(\sin\theta(t, \nu') - \theta(t, \nu')\cos\bar\theta)\right| dt + (\nu'-\nu)\cos\bar\theta,\end{aligned} \tag{5.20}$$

which, using (5.6) for the function $g(t) = \sin\theta(t, \nu') - \theta(t, \nu')\cos\bar\theta$, yields (5.15). □

The bounds (5.14) and (5.15) take advantage, through the total variation of the $\theta(\cdot, \nu')$ and $\sin\theta(\cdot, \nu') + \theta(\cdot, \nu')\cos\bar\theta$ functions, of much of the information contained in the shape of the deflection function $\theta(\cdot, \nu')$.

In particular, the lower bound (5.15) for the numerator of $\rho_G(\nu, \nu')$ increases when the variation of the deflection increases, which shows that "oscillating paths" have more chance to have positive global radii of curvature.

If we retain only from the shape of $\theta(\cdot, \nu')$ its maximum value $\bar{\theta}$ on $[\nu, \nu']$, we obtain the following corollary.

COROLLARY 5.4. *Let $\nu, \nu' \in [0, \delta(P)]$ be given. Then*

$$\bar{\theta} \leqq \int_{\min(\nu,\nu')}^{\max(\nu,\nu')} \frac{d\nu}{\rho(\nu)} \leqq \frac{|\nu' - \nu|}{\bar{R}}, \tag{5.22}$$

$$\operatorname{sgn}(\nu' - \nu)\langle X' - X, v'\rangle \geqq \bar{R} \sin \bar{\theta} + (|\nu' - \nu| - \bar{R} \cdot \bar{\theta}) \cos \bar{\theta}, \tag{5.23}$$

where $\bar{R}$, $\bar{\theta}$ are defined in (5.16) *and* (5.17).

Proof. Suppose $\nu \leqq \nu'$, for instance, and define $\bar{\nu} \in [\nu, \nu']$ by

$$\theta(\bar{\nu}, \nu') = \bar{\theta} = \sup \theta(t, \nu') \qquad t \in [\nu, \nu'].$$

Then (5.22) and (5.23) result immediately from (5.14), (5.15), and (5.4) with $N = 1$, $t_0 = \bar{\nu}$, $t_1 = \nu'$. □

We can put together formula (4.4) giving $\rho_G(\nu, \nu')$ and Corollary 5.4 to obtain the sought lower bound on $\rho_G(\nu, \nu')$.

PROPOSITION 5.5. *Let $P \in \mathcal{P}$, $\nu, \nu' \in [0, \delta(P)]$, $\nu \neq \nu'$, be given, and $\bar{R}$, $\bar{\theta}$ be defined by* (5.16) *and* (5.17).

(i) *If*

$$0 \leqq \bar{\theta} \leqq \pi/2, \tag{5.24}$$

then

$$\rho_G(\nu, \nu') \geqq \bar{R} + (|\nu' - \nu| - \bar{R} \cdot \bar{\theta}) \operatorname{cotan} \bar{\theta} \tag{5.25}$$

and the right-hand side of (5.25) *is strictly positive as soon as*

$$\bar{R} > 0 \qquad (\textit{independently of } |\nu' - \nu|!). \tag{5.26}$$

(ii) *If*

$$\frac{\pi}{2} \leqq \bar{\theta} \leqq \pi \tag{5.27}$$

then

$$\rho_G(\nu, \nu') \geqq \bar{R} \sin \bar{\theta} + (|\nu' - \nu| - \bar{R} \cdot \bar{\theta}) \cos \bar{\theta} \tag{5.28}$$

and the right hand side of (5.28) *is strictly positive as soon as the size × curvature condition*

$$|\nu' - \nu| / \bar{R} < \bar{\theta} - \tan \bar{\theta} \tag{5.29}$$

is satisfied (*note that, because of* (5.22), *condition* (5.29) *can be satisfied only of $\bar{\theta} < \pi$!*).

Proof. First we prove (i). As $\bar{\theta} \leqq \pi/2$ we have $\cos \bar{\theta} \geqq 0$, which shows by Corollary 5.4 that

$$\operatorname{sgn}(\nu' - \nu)\langle X' - X, v'\rangle \geqq 0, \tag{5.30}$$

and by definition of θ we have

$$\theta(\nu, \nu') \leqq \bar{\theta} \leqq \pi/2, \tag{5.31}$$

which shows that

$$\cos \theta(\nu, \nu') = \langle v, v'\rangle \geqq 0 \tag{5.32}$$

and

$$0 \leqq \sin \theta(\nu, \nu') \leqq \sin \bar{\theta}. \tag{5.33}$$

Using (5.30) and (5.32) in formula (4.4) for ρ_G then yields

$$\rho_G(\nu, \nu') = \frac{\operatorname{sgn}(\nu' - \nu)\langle X' - X, v'\rangle}{\sin \theta(\nu, \nu')}, \tag{5.34}$$

and, using (5.33),

$$\rho_G(\nu, \nu') \geqq \frac{\operatorname{sgn}(\nu' - \nu)\langle X' - X, v'\rangle}{\sin \bar{\theta}}, \tag{5.35}$$

which yields the sought result (5.25) using the lower bound (5.23) for the numerator.

Now we prove (ii). As $\pi/2 \leqq \theta < \pi$, we have no information on the sign of $\cos \theta(\nu, \nu') = \langle v, v'\rangle$. However, we see from (4.4) that, whatever the signs of $\operatorname{sgn}(\nu' - \nu) \times \langle X' - X, v'\rangle$ and $\langle v, v'\rangle$ may be, we always have

$$\rho_G(\nu, \nu') \geqq \operatorname{sgn}(\nu' - \nu)\langle X' - X, v'\rangle, \tag{5.36}$$

the equality holding when $\operatorname{sgn}(\nu' - \nu)\langle X' - X, v'\rangle \geqq 0$ and $\langle v, v'\rangle \leqq 0$. Then the sought lower bound (5.28) results immediately from (5.36) and formula (5.23) of Corollary 5.4, and the size × curvature condition (5.29) results immediately from (5.28) by noting that $\cos \bar{\theta} \leqq 0$. □

Formula (5.25) also gives us a more precise insight into the way $\rho_G(\nu, \nu')$ approaches $\rho(\nu)$ when $\nu' \to \nu$: if, for example, $t \to \rho(t)$ is an increasing function, then

$$\begin{aligned} &\rho_G(\nu, \nu') \geqq \rho(\nu) \quad \text{when } \nu' \geqq \nu, \\ &\rho_G(\nu, \nu') \geqq \rho(\nu') \quad \text{when } \nu' \leqq \nu \end{aligned} \tag{5.37}$$

as soon as ν' and ν are close enough so that the maximum deflection between v and v' is smaller than $\pi/2$.

We now deduce from Proposition 5.5 a lower bound for $R_G(P) = \inf_{\nu, \nu' \in [0, \delta(P)]} \rho_G(\nu, \nu')$, which is our second main result.

THEOREM 5.6. *Let a path $P \in \mathcal{P}$ be given. Then its maximum deflection $\Theta(P)$ (defined in (5.2)) is related to its length $\delta(P)$ and its smallest radius of curvature $R(P)$ by*

$$\Theta(P) \leqq \int_0^{\delta(P)} \frac{d\nu}{\rho(\nu)} \leqq \frac{\delta(P)}{R(P)}, \tag{5.38}$$

and the following lower bounds for $R_G(P)$ defined in (4.17) hold:

(i) *Low deflection paths. If*

$$0 \leqq \Theta(P) \leqq \pi/2 \tag{5.39}$$

then

$$R_G(P) = R(P) \tag{5.40}$$

independently of the length $\delta(P)$ of the path.

(ii) *Large deflection paths. If*

$$\pi/2 \leqq \Theta(P) \leqq \pi \tag{5.41}$$

then

$$R_G(P) \geqq R(P) \sin \Theta(P) + \{\delta(P) - R(P)\Theta(P)\} \cos \Theta(P), \tag{5.42}$$

which is strictly positive as soon as the length $\delta(P)$ of the path satisfies the size × curvature condition:

$$\delta(P)/R(P)<\Theta(P)-\tan\Theta(P), \tag{5.43}$$

(*once again, this condition can be satisfied only if* $\Theta(P)<\pi$!).

Proof. Note first that (5.38) is obtained immediately from (5.22) of Corollary 5.4 by taking $\nu=\nu_0$, $\nu'=\nu_1$, where $\nu_0, \nu_1\in[0, \delta(P)]$ are such that $\theta(\nu_0, \nu_1)=\Theta(P)$, and by noting that

$$\int_{\nu_0}^{\nu_1}\frac{d\nu}{\rho(\nu)}\leqq\int_0^{\delta(P)}\frac{d\nu}{\rho(\nu)}.$$

We now prove the results of (i). If $\Theta(P)\leqq\pi/2$, we obviously have for any $\nu, \nu'\in[0, \delta(P)]$,

$$0\leqq\bar\theta\leqq\Theta(P)\leqq\pi/2$$

so that part (i) of Proposition 5.5 applies. Hence

$$\rho_G(\nu, \nu')\geqq\bar R,$$

which implies, as $\bar R\geqq R(P)$, that

$$R_G(P)\geqq R(P).$$

But we know from Proposition 4.7 that

$$R_G(P)\leqq R(P),$$

which ends the proof of (5.40).

We now prove the results of (ii). As we now know only that $\pi/2\leqq\Theta(P)\leqq\pi$, two cases may happen for given $\nu, \nu'\in[0, \delta(P)]$: either $0\leqq\bar\theta\leqq\pi/2\leqq\Theta(P)$, and we find, as above,

$$\rho_G(\nu, \nu')\geqq R(P) \tag{5.44}$$

or $\pi/2\leqq\bar\theta\leqq\Theta(P)\leqq\pi$ and then we find from part (ii) of Proposition 5.5 that

$$\rho_G(\nu, \nu')\geqq\bar R\sin\bar\theta+(|\nu'-\nu|-\bar R\cdot\bar\theta)\cos\bar\theta.$$

However, since $\cos\bar\theta\leqq0$, this implies, as $|\nu'-\nu|\leqq\delta(P)$ and $\bar R\geqq R(P)$, that

$$\rho_G(\nu, \nu')\geqq R(P)\sin\bar\theta+(\delta(P)-R(P)\bar\theta)\cos\bar\theta.$$

However, the mapping $\alpha\to R(P)\sin\alpha+(\delta(P)-R(P)\alpha)\cos\alpha$ is decreasing over the $[0, \Theta(P)]$ interval, as its derivative is the $\alpha\to-\sin\alpha(\delta(P)-R(P)\alpha)$ function, which is negative over the $[0, \Theta(P)]$ interval because of (5.38). Hence we obtain, as $\bar\theta\leqq\Theta(P)$,

$$\rho_G(\nu, \nu')\geqq R(P)\sin\Theta(P)+(\delta(P)-R(P)\Theta(P))\cos\Theta(P). \tag{5.45}$$

When taking the minimum over all $\nu, \nu'\in[0, \delta(P)]$, we obtain (5.42) because the right-hand side of (5.45) is smaller than the right-hand side of (5.44).

Finally, (5.43) is obtained immediately by writing that the right-hand side of (5.42) is strictly positive. □

We have illustrated Theorem 5.6 in Figs. 5.1 and 5.2. For a path P with given smallest radius of curvature $R(P)$, Fig. 5.1 shows the *domain* of deflections $\Theta(P)$ and length $\delta(P)$, which are *recognized by Theorem* 5.6 *as associated to a strictly quasiconvex* path P. It is delimited by the following two curves:

$$\delta_{\max}(P)=\begin{cases}+\infty & \text{if } 0\leqq\Theta(P)\leqq\pi/2,\\ R(P)\{\Theta(P)-\tan\Theta(P)\} & \text{if } \pi/2<\Theta(P)\leqq\pi,\end{cases}$$

$$\delta_{\min}=R(P)\Theta(P).$$

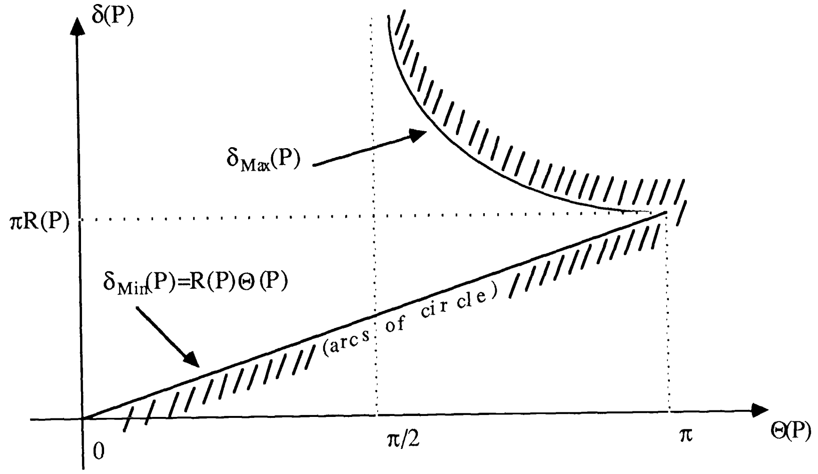


FIG. 5.1. *The domain of deflections* $\Theta(P)$ *and lengths* $\delta(P)$ *that are recognized as strictly quasiconvex by Theorem* 5.6 *for a path with a given* $R(P)$.

Note that (5.38) implies that the bottom boundary $\delta_{\min}(P)$ is a hard one, in the sense that there do not exist paths whose representative point is below that line. On the contrary, the upper boundary is soft, as there exist many paths having their representative points above that boundary, and that may or may not be strictly quasiconvex.

Figure 5.2 represents, for all paths recognized as strictly quasiconvex by Theorem 5.6, the domain of deflections $\Theta(P)$ and lower bounds R_G to $R_G(P)$ given by Theorem 5.6, namely,

$$R_G=\begin{cases} R(P) & \text{if } 0\leqq\Theta(P)\leqq\pi/2,\\ R(P)\sin\Theta(P)+\{\delta(P)-R(P)\Theta(P)\}\cos\Theta(P) & \text{if } \pi/2\leqq\Theta(P)\leqq\pi.\end{cases}$$

Note that here the upper boundary, which corresponds to paths that are arcs of circles, is not a hard boundary for the actual global radius of curvature $R_G(P)$: we may find paths with a deflection $\Theta(P)$ as close as desired from π but with a $R_G(P)=R(P)$ global radius of curvature!

We note also that the bounds given in Theorem 5.6 are sharp; when P is an arc of a circle, the inequalities in (5.38) and (5.42) become equalities, and the theorem recognizes exactly all strictly quasiconvex arcs of circles, as the equality in (5.38) implies that (5.43) is satisfied as soon as $\delta(P)/R(P)<\pi$. But the upper bound (5.43) for the admissible length $\delta(P)$ of a path with a maximum deflection larger than $\pi/2$ is also sharp, as may be seen from Fig. 5.3: given a maximum deflection $\pi/2\leqq\Theta(P)<\pi$, the worst thing to do (up to a change of scale, of course) is

—first go straight ahead with a length $|\tan\Theta(P)|$

—then turn an angle $\Theta(P)$ using an arc of a circle of radius one.

The resulting path has a length $\Theta(P)-\tan\Theta(P)$, and a zero $R_G(P)$ as obviously $\rho_G(0,\delta(P))=0$!

Of course, we may use Theorem 5.6 to construct various sufficient conditions for the strict quasiconvexity of a set, as we check easily from Theorem 4.12 that the only thing we must do is to find a lower bound for $R_G(P)$, $P\in\mathcal{P}_M$.

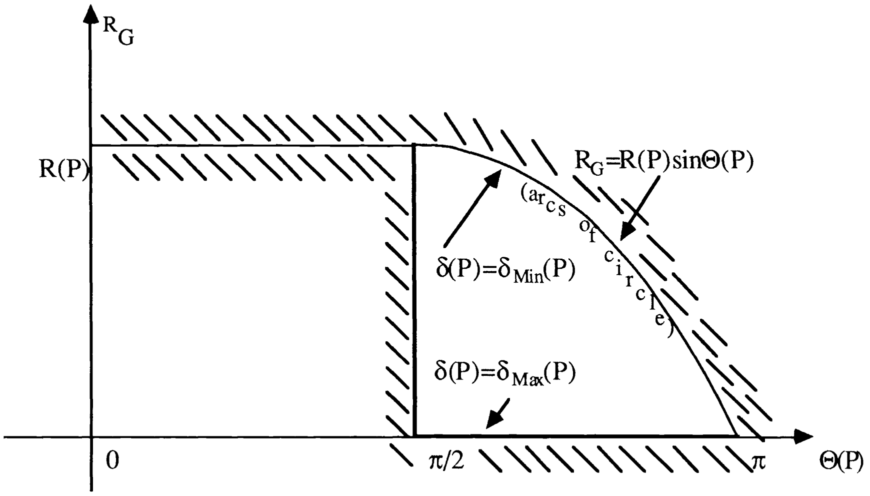


FIG. 5.2. *The domain of deflections* $\Theta(P)$ *and lower bounds* R_G *to* $R_G(P)$ *given by Theorem* 5.6 *for paths recognized by this theorem as in Fig.* 5.1.

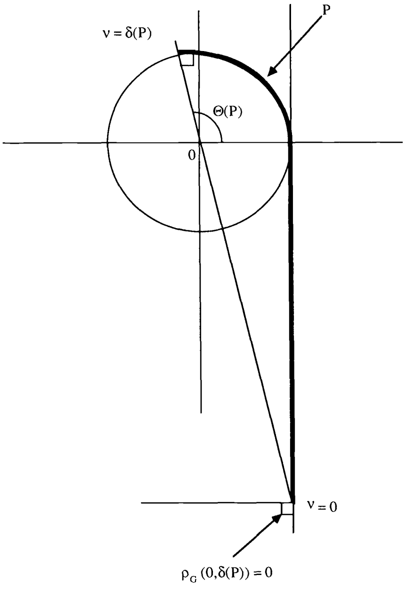


FIG. 5.3. *Illustration of the sharpness of the upper bound for the length of paths with a maximum deflection larger than* $\pi/2$.

We state here one such condition, which uses the following lower and upper bounds for the attributes of the set $(D, \mathcal{P})$:

Lower bound R to radii of curvature along paths:

$$R(P) \geqq R \quad \forall P \in \mathcal{P}_M, \tag{5.46}$$

Upper bound Δ *to path length*:

$$\delta(P) \leqq \Delta \quad \forall P \in \mathcal{P}_M, \tag{5.47}$$

Upper bound Θ to path maximal deflection:

$$\Theta(P) \leqq \Theta \quad \forall P \in \mathcal{P}_M. \tag{5.48}$$

Note that such an upper bound can be obtained by either formula

$$\Theta = \sup_{P \in \mathcal{P}_M} \int_0^{\delta(P)} \frac{d\nu}{\rho(\nu)} \quad \text{or} \quad \Theta = \sup_{P \in \mathcal{P}_M} \frac{\delta(P)}{R(P)}. \tag{5.49}$$

THEOREM 5.7. (Θ-size × curvature condition). *Let $(D, \mathcal{P})$ be given such that*

$$R > 0. \tag{5.50}$$

(i) *If*

$$0 \leqq \Theta \leqq \pi/2, \tag{5.51}$$

then

$$R_G(D) \geqq R > 0, \tag{5.52}$$

and $(D, \mathcal{P})$ is strictly quasiconvex for a cylindrical neighborhood of size at least R.

(ii) *If*

$$\pi/2 \leqq \Theta < \pi, \tag{5.53}$$

then

$$R_G(D) \geqq R \sin \Theta + (\Delta - R\Theta) \cos \Theta, \tag{5.54}$$

and $(D, \mathcal{P})$ is strictly quasiconvex, with a cylindrical neighborhood of the size of the right-hand side of (5.54) *at least, as soon as the maximum length of paths satisfy*

$$\Delta / R < \Theta - \tan \Theta. \tag{5.55}$$

6. One application to nonlinear least squares. We consider the problem

$$\text{find } x \in C \text{ such that } J(x) = \|\varphi(x) - z\|_F^2 = \min \text{ over } C, \tag{6.1}$$

where

$$\begin{aligned} &E \text{ is a Banach space, } F \text{ is a Hilbert space,} \\ &C \subset E \text{ is a closed convex set,} \\ &\varphi : C \to F \text{ is a } C^2\text{-mapping,} \\ &z \in F \text{ is a given point (data).} \end{aligned} \tag{6.2}$$

Because φ is of class C^2 there exists β such that

$$\|\varphi''(x)(y, y)\|_F \leqq \beta \|y\|^2 \quad \forall x \in C, \quad \forall y \in E, \tag{6.3}$$

and we suppose that the derivative $\varphi'(x)$ is uniformly invertible, i.e., that there exist α_m and α_M such that

$$\alpha_m \|y\|_E \leqq \|\varphi'(x) \cdot y\|_F \leqq \alpha_M \|y\|_E \quad \forall x \in C, \quad \forall y \in E. \tag{6.4}$$

We may then apply the results of § 5 to the set $D = \varphi(C)$ equipped with the paths

$$\mathcal{P} = \{\varphi([x, y]), x, y \in C\} \tag{6.5}$$

for which the subset

$$\mathcal{P}_M = \{\varphi([x, y]), x, y \in \partial C\} \tag{6.6}$$

is obviously a subfamily of maximal paths.

We easily check (cf. [6]) that, given a path $P=\varphi([x,y])$, x, $y\in C$, the arc length ν and radius of curvature ρ satisfy

$$\nu(t)=\int_0^t \|\varphi'(x_t)(y-x)\|_F\,dt \geqq \alpha_m t\|y-x\|_E, \tag{6.7}$$

$$\rho(\nu)\geqq \frac{\|\varphi'(x_t)(y-x)\|_F^2}{\|\varphi''(x_t)(y-x,y-x)\|_F}\geqq \frac{\alpha_m^2}{\beta}, \tag{6.8}$$

where

$$x_t=(1-t)x+ty \quad \forall t\in[0,1]. \tag{6.9}$$

From (6.7) and (6.8) we find that

$$\int_0^{\delta(P)} \frac{d\nu}{\rho(\nu)}\leqq \int_0^1 \frac{\|\varphi''(x_t)(y-x,y-x)\|_F}{\|\varphi'(x_t)(y-x)\|_F}\,dt \leqq \frac{\beta}{\alpha_m}\|y-x\|_E. \tag{6.10}$$

These formulas give us the following lower and upper bounds for the attributes of the paths of $D=\varphi(C)$:

$$\begin{aligned} &R=\alpha_m^2/\beta && \text{(lower bound to radii of curvature),}\\ &\Delta=\alpha_M \operatorname{diam} C && \text{(upper bound to path length),}\\ &\Theta=(\beta/\alpha_m)\operatorname{diam} C && \text{(upper bound to path deflection).}\end{aligned} \tag{6.11}$$

Using Theorem 5.7 we then obtain the following theorem.

THEOREM 6.1. *Let* (6.2)-(6.4) *hold, and let* R, Δ, Θ *be defined by* (6.11), *and define* R_G *by*

$$R_G=\begin{cases} R & \text{if } 0\leqq\Theta\leqq\pi/2,\\ R\sin\Theta+(\alpha_M-\alpha_m)\operatorname{diam} C\cos\theta & \text{if } \pi/2\leqq\Theta\leqq\pi.\end{cases} \tag{6.12}$$

If

$$\Theta=\frac{\beta}{\alpha_m}\operatorname{diam} C\leqq\frac{\pi}{2}, \tag{6.13}$$

or

$$\frac{\pi}{2}\leqq\Theta=\frac{\beta}{\alpha_m}\operatorname{diam} C<\pi \quad\text{and}\quad \left(\frac{\alpha_M}{\alpha_m}-1\right)\Theta<|\tan\Theta|, \tag{6.14}$$

then, for any z *such that*

$$d(z,\varphi(C))<R_G, \tag{6.15}$$

the nonlinear least squares problem (6.1) *has the following properties*:

(6.16) *existence of a unique global minimum* $\hat{x}$,

(6.17) *no local minimum,*

(6.18) *any minimizing sequence is a Cauchy sequence in* E *converging toward* $\hat{x}$,

(6.19) *Lipschitz-continuity of the* $z\to\hat{x}$ *mapping*:

$$\alpha_m\|\hat{X}_1-\hat{X}_2\|_E\leqq\left(1-\frac{d}{R}\right)^{-1}\|z_1-z_2\|_F \text{ as soon as}$$

$$\|z_1-z_2\|_F+\max_{j=1,2} d(z_j,\varphi(C))\leqq d<R_G. \tag{6.20}$$

Note that, under the same hypothesis, the γ-size curvature condition would read

$$\frac{\beta}{\alpha_m}\operatorname{diam} C < 2\sqrt{2} \tag{6.21}$$

and imply (6.16)-(6.19) but not (6.17), and (6.20) would hold only for $d < \gamma(D)$, which is usually much smaller than R_G.

REFERENCES

[1] T. Abatzoglou, *Unique best approximation from a C^2-manifold in Hilbert space*, Pacific J. Math., 87 (1980), pp. 233-244.

[2] J. P. Aubin, *Mathematical Methods of Game and Economic Theory*, North-Holland, Amsterdam, 1979.

[3] G. Chavent, *Local stability of the output least squares parameter estimation techniques*, Mat. Apl. Comput., 2 (1983), pp. 3-22.

[4] ———, *On the uniqueness of local minima for general abstract non-linear least squares problems*, Inverse Problems, 4 (1988), pp. 417-433.

[5] ———, *A new sufficient condition for the wellposedness of non-linear least-squares problems arising in identification and control*, in Analysis and Optimization of Systems, Lecture Notes in Control and Inform. Sci., Vol. 144, A. Bensoussan and J. L. Lions, eds., Springer-Verlag, New York, Berlin, 1990, pp. 452-463.

[6] ———, *Quasiconvex sets and the size × curvature condition*: *Application to non-linear inversion*, J. Appl. Math. Optim., to appear.

[7] J. R. Rice, *Nonlinear approximation* II. *Curvature in Minkowski geometry and local uniqueness*, Trans. Amer. Math. Soc., 128 (1967), pp. 437-459.

[8] W. Rudin, *Real and Complex Analysis*, McGraw-Hill, New York, 1987.

SIAM J. CONTROL AND OPTIMIZATION.
Vol. 29, No. 6, pp. 1373–1393, November 1991

H_∞ CONTROL WITH TRANSIENTS*

PRAMOD P. KHARGONEKAR[†], KRISHAN M. NAGPAL[‡], AND KAMESHWAR R. POOLLA[§]

Abstract. In H_∞ (or uniformly optimal) control problems, it is usually assumed that the system initial conditions are zero. In this paper, an H_∞-like control problem that incorporates uncertainty in initial conditions is formulated. This is done by defining a worst-case performance measure. Both finite and infinite horizon problems are considered. Necessary and sufficient conditions are derived for the existence of controllers that yield a closed-loop system for which the above-mentioned performance measure is less than a prespecified value. State-space formulae for the controllers are also presented.

Key words. H_∞ control theory, algebraic and differential Riccati equations, optimal control

AMS(MOS) subject classifications. 93B50, 93C35, 93C05, 49A40

1. Introduction.

Zames introduced the problem of H_∞ optimal control in his pioneering paper [18]. The essential idea was to design a controller to optimize the closed-loop system performance for the worst exogenous input. The expository book by Francis [6] presents a lucid account of the early developments in H_∞ control theory.

A significant new development in H_∞ control theory in the last two years has been the introduction of state-space methods. This has led to a rather transparent solution to the standard problem of H_∞ control theory. See Doyle et al. [5], Khargonekar [9], and the references cited there for the state-space approach to H_∞ control theory.

The H_∞ norm of a system can be defined in many different but equivalent ways. However, it is always (at least, implicitly) assumed that the initial condition of the system is zero. Thus, in most of the H_∞ control theory literature, it is assumed that the plant initial conditions are zero. There are a few exceptions. For example, Nagpal and Khargonekar [12] have considered an H_∞ type of estimation problem with nonzero initial conditions. In this paper, our principal aim is to extend the basic ideas and the recent results from the state-space approach to H_∞ control theory taking initial conditions into account explicitly. We consider this as the key conceptual contribution of the present paper. In recent independent parallel work, Didinsky and Basar [4] consider a minimax design problem for discrete-time systems with nonzero initial states. However, their problem formulation, as well as the results, bear little resemblance to our work.

In §2 we formulate an H_∞-type optimal control problem that incorporates initial conditions. This is done by introducing a new performance measure that is essentially the worst-case norm of the regulated outputs over all exogenous signals and initial conditions. We define this performance measure for both finite time and infinite time

* Received by the editors March 5, 1990; accepted for publication (in revised form) September 10, 1990. This research was supported in part by National Science Foundation grants ECS-9096109 and ECS-8957461, U.S. Air Force Office of Scientific Research contract AFOSR-90-0053, U.S. Army Research Office grant DAAL03-90-G-0008, and grants from Honeywell, McDonnell-Douglas, and Rockwell International. The results of this paper were presented at the Dutch System and Control Theory Network summer school on H_∞ Control and Robust Stabilization, Schiermonnikoog, the Netherlands, June 1990, and at the 1990 IEEE Conference on Decision and Control, December 1990.

[†] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109-2122.

[‡] Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, Michigan 48109-2122. Present address, Coordinated Sciences Laboratory, University of Illinois, Urbana, Illinois, 61801.

[§] Coordinated Sciences Laboratory, University of Illinois, Urbana, Illinois, 61801.

horizons. For finite horizon problems we also allow for a penalty on the terminal state. This enables us to incorporate trade-offs between the norm of the controlled output and the size of the terminal state.

In §3 we state the main results of this paper. Here we present necessary and sufficient conditions for the existence of a linear controller such that the above-mentioned performance measure of the closed-loop system is less than a prespecified number. These necessary and sufficient conditions are given in terms of existence and properties of solutions to certain algebraic and differential Riccati equations. In the event that these necessary conditions are met, we provide explicit formulae for controllers that yield the prespecified performance. Our results in this paper may be regarded as the H_∞ analogue of the *nonstationary* linear quadratic Gaussian (LQG) control theory results. The results for infinite time horizon problems are natural generalizations of the results of Doyle et al. [5] for the situation of nonzero initial states, while those for finite time horizon problems are natural generalizations of the results of Tadmor [15] and Limebeer et al. [10]. In §4 we give proofs of the main results.

In this paper, we restrict our attention to finite-dimensional linear time-invariant plants. It should be noted that even though the plant is linear time-invariant, it is necessary to consider time-varying controllers since the natural (central) solutions even for linear time-invariant plants happen to be linear *time-varying.* This situation is analogous to the the finite horizon linear-quadratic regulator and Kalman filtering theory. Recall that for finite horizon linear-quadratic optimal control problems for linear time-invariant plants, the optimal controller turns out to be linear time-varying. Similarly, in the Kalman filtering problem, if the initial state covariance does not match the steady-state covariance, the Kalman filter is also linear time-varying. It is in this sense in which we regard our results in this paper as the H_∞ analogue of the nonstationary LQG control theory results. Results for finite horizon problems can be trivially extended to linear time-varying plants. Extensions to time-varying plants of the results for the infinite time horizon case are technically much more difficult but are possible along the lines of recent work by Tadmor [16] and Ravi, Nagpal, and Khargonekar [13]. These extensions are left for future research.

2. Worst-case H_∞-type performance measures with nonzero initial conditions.

Consider the finite-dimensional linear time-invariant system Σ:

$$\begin{aligned} \frac{dx}{dt} &= Ax + B_1w + B_2u, \\ z &= C_1x + D_{11}w + D_{12}u, \\ y &= C_2x + D_{21}w + D_{22}u. \end{aligned} \tag{1}$$

Here x, w, u, z, and y denote, respectively, the state, exogenous input, control input, regulated output, and measured output. It is assumed that the initial state $x(0)$ is possibly nonzero and *unknown.*

The control problem that we wish to address is that of designing a controller that internally stabilizes the closed-loop system and reduces z uniformly for all w and $x(0)$. More specifically, let K be a finite-dimensional linear (possibly time-varying) controller given by the system equations

$$\begin{aligned} \frac{d\xi}{dt} &= F(t)\xi(t) + G(t)y(t), \\ u(t) &= H(t)\xi(t) + J(t)y(t). \end{aligned} \tag{2}$$

Throughout this paper, we assume that the controllers are linear, finite-dimensional, and time-varying with continuous and bounded state-space realizations. (A time function $f(t)$ is called bounded if and only if there exists $M > 0$ such that for all $t \geq 0$ $\|f(t)\| < M$.) Let Σ_{cl} denote the resulting closed-loop system. The closed-loop system is called *well posed* if and only if $(I - JD_{22})^{-1}$ is bounded. The closed-loop system is called *internally stable* if and only if it is well posed and the unforced closed-loop system (i.e., $w = 0$,) with states x, ξ is exponentially stable. In finite horizon problems, a controller is called *admissible* if and only if it yields a well-posed feedback system. In infinite horizon problems, a controller is called *admissible* if and only if it yields an internally stable feedback system. For a fixed time $T > 0$, a symmetric positive-semidefinite matrix S, and a symmetric positive-definite matrix R, define the worst-case closed-loop performace measure as

$$J(\Sigma_{cl}, R, S, T) = \sup \left\{ \left[\frac{\|z\|_T^2 + x'(T)Sx(T)}{\|w\|_T^2 + x_0'Rx_0} \right]^{1/2} \right\}, \tag{3}$$

where $\xi(0) = 0$, and the supremum is taken over all $x(0) = x_0 \in \mathbf{R}^n$, $w \in L_2[0, T]$, $\|w\|_T^2 + x_0'Rx_0 \neq 0$. In this definition, T is allowed to be ∞ in which case $S := 0$ and the supremum on the right-hand side is taken over all $w \in L_2[0, \infty)$. Here $\|f\|_T = \left[\int_0^T \|f(t)\|^2 dt\right]^{1/2}$.

The performance measure $J(\Sigma_{cl}, R, S, T)$ can be regarded as the induced norm of the linear operator generated by the closed-loop system Σ_{cl}, which maps the pair $(x(0), w)$ to $(x(T), z)$. More explictly, consider the linear operator

$$\Gamma : \mathbf{R}^n \oplus L_2[0, T] \to \mathbf{R}^n \oplus L_2[0, T] : \ (x(0), w) \to (x(T), z).$$

Define the inner product on the domain of Γ to be

$$\langle (x_1, w_1), (x_2, w_2) \rangle := x_1'Rx_2 + \langle w_1, w_2 \rangle_{L_2[0,T]}$$

and the inner product on the co-domain as

$$\langle (x_1, z_1), (x_2, z_2) \rangle := x_1'Sx_2 + \langle z_1, z_2 \rangle_{L_2[0,T]},$$

where $<>_{L_2[0,T]}$ is the usual inner product in $L_2[0, T]$. These inner products induce corresponding (semi)norms in the domain and the range of Γ. Then the performance measure J is the induced operator norm of Γ. A similar interpretation can be given for the infinite horizon case.

We can now state the control problems considered in this paper. Given a real number $\gamma > 0$:

(i) Infinite Horizon Problem: Does there exist, and if so find, an internally stabilizing bounded linear time-varying controller such that

$$J(\Sigma_{cl}, R, 0, \infty) < \gamma?$$

(ii) Finite Horizon Problem: Does there exist, and if so find, a bounded linear time-varying controller such that

$$J(\Sigma_{cl}, R, S, T) < \gamma?$$

In the remainder of this section, we will present some results that serve in developing an understanding of the performance measure J. These results also elucidate the relationship between the cost functional J and the H_∞ norm of a system.

Consider the finite-dimensional linear (possibly time-varying) system Σ_a

$$\frac{dx}{dt} = Ax + Bw, \qquad z = Cx \tag{4}$$

and define $J(\Sigma_a, R, S, T)$ as above. We summarize some of the simpler properties of $J(\Sigma_a, R, S, T)$ in the following lemma.

LEMMA 1.1. *Let Σ_a be as above.*

(a) *If $R_1 \geq R_2$ then $J(\Sigma_a, R_1, S, T) \leq J(\Sigma_a, R_2, S, T)$.*

(b) *If $S_1 \geq S_2$, then $J(\Sigma_a, R, S_1, T) \geq J(\Sigma_a, R, S_2, T)$.*

(c) *If $T_1 \geq T_2$, then $J(\Sigma_a, R, 0, T_1) \geq J(\Sigma_a, R, 0, T_2)$.*

(d) *Suppose that Σ_a is time-invariant and asymptotically stable, then*

$$\lim_{\rho, T \to \infty} J(\Sigma_a, \rho I, 0, T) = \|T_{zw}\|_\infty := \sup\{\bar{\sigma}(T_{zw}(s)) : \operatorname{Re}(s) \geq 0\}, \tag{5}$$

where $T_{zw} := C(sI - A)^{-1}B$ denotes the system transfer function and, moreover,

$$J(\Sigma_a, R, 0, \infty) \geq \|T_{zw}\|_\infty. \tag{6}$$

Thus, $J(\Sigma_{cl}, R, S, T)$ is a generalization of the more familiar concept of the H_∞ norm of a system accommodating the possibility of nonzero initial conditions. The weighting matrix R is a measure of the relative importance of the uncertainty in initial conditions vis-à-vis the uncertainty in the exogenous signals w. A "smaller" choice of R reflects greater uncertainty in the initial condition. This connection is best illuminated by observing that as the smallest eigenvalue $\lambda_{\min}(R)$ of R approaches ∞, the unit ball in $\mathbf{R}^n \oplus L_2$ defined by

$$B_{\mathbf{R}^n \oplus L_2} := \left\{(x_0, w) \in \mathbf{R}^n \oplus L_2 : x_0' R x_0 + \|w\|^2 \leq 1\right\}$$

tends to $(0, B_{L_2})$, where $B_{L_2} := \{w \in L_2 : \|w\|^2 \leq 1\}$.

We now describe how the performance measure J can be computed for a given system. These results may be viewed as natural generalizations of existing work on the computation of the H_∞ norm of a linear time-invariant system (see, for example, Anderson [1], Willems [17], and Boyd, Balakrishnan, and Kabamba [2]).

THEOREM 1.2. *Consider the linear (possibly time-varying) system Σ_a as in (4) above. Let R, S be given symmetric matrices such that S is positive semidefinite and R is positive definite. Then the following are equivalent:*

(a) $J(\Sigma_a, R, S, T) < \gamma$.

(b) *There exists a symmetric matrix function $P(t), t \in [0, T]$ such that*

$$-\dot{P}(t) = A'(t)P(t) + P(t)A(t) + \frac{1}{\gamma^2}P(t)B(t)B'(t)P(t) + C'(t)C(t), P(T) = S \tag{7}$$

and $P(0) < \gamma^2 R$.

(c) *There exists a symmetric matrix function $Q(t) > 0, t \in [0, T]$ such that*

$$\dot{Q}(t) = A(t)Q(t) + Q(t)A'(t) + \frac{1}{\gamma^2}Q(t)C'(t)C(t)Q(t) + B(t)B'(t), Q(0) = R^{-1} \tag{8}$$

and $\gamma^2 Q^{-1}(T) > S$.

THEOREM 1.3. *Let* Σ_a *in* (4) *be a linear time-invariant, asymptotically stable system. Let* R *be a given symmetric positive-definite matrix. Then the following are equivalent:*

(a) $J(\Sigma_a, R, 0, \infty) < \gamma$.

(b) *There exists a symmetric matrix* P *such that*

$$0 = A'P + PA + (1/\gamma^2)PBB'P + C'C,$$

$(A + (1/\gamma^2)BB'P)$ *is asymptotically stable, and* $P < \gamma^2 R$.

(c) *There exists a symmetric* $Q(t)$ *that satisfies the Riccati differential eqution* (8) *for all* $t \geq 0$ *and is such that the autonomous system* $\dot{q}(t) = [A + (1/\gamma^2)Q(t)C'C]q(t)$ *is exponentially stable. Moreover,* $\lim_{t\to\infty} Q(t)$ *exists and equals* Q_∞*, where* Q_∞ *is the unique symmetric matrix with the following properties:*

$$AQ_\infty + Q_\infty A' + \frac{1}{\gamma^2}Q_\infty C'CQ_\infty + BB' = 0, \tag{9}$$

and $A + (1/\gamma^2)QC'C$ *is asymptotically stable.*

Proofs of Theorems 2.2 *and* 2.3. Equivalence of (a) and (b). We will first prove that (a) ⇒ (b) in Theorem 2.3. From Lemma 2.1, it follows that $J(\Sigma_a, R, 0, \infty) \geq \|C(sI - A)^{-1}B\|_\infty$. It follows from Anderson [1], Willems [17], and Boyd, Balakrishnan, and Kabamba [2] that $\|C(sI - A)^{-1}B\|_\infty < \gamma$, if and only if there exists a symmteric matrix P such that

$$0 = A'P + PA + \frac{1}{\gamma^2}PBB'P + C'C,$$

and $(A + (1/\gamma^2)BB'P)$ is asymptotically stable. To complete the necessity it remains to be shown that $P < \gamma^2 R$. Suppose that this is not the case. Then there exists a nonzero x_0 such that $x_0'(P - \gamma^2 R)x_0 \geq 0$. Straightforward algebra reveals that for system (4),

$$\frac{d(x'Px)}{dt} = \gamma^2 w'w - x'C'Cx - \left[\gamma w - \frac{1}{\gamma}B'Px\right]' \left[\gamma w - \frac{1}{\gamma}B'Px\right].$$

Now set $w = (1/\gamma^2)B'Px, x(0) = x_0$. Using stability of $(A + (1/\gamma^2)BB'P)$ and integrating the above equation from 0 to ∞ along the trajectory of Σ_a, we obtain

$$\gamma^2 x_0' R x_0 + \gamma^2\|w\|^2 - \|z\|^2 = \gamma^2 x_0' R x_0 - x_0' P x_0 \leq 0,$$

which contradicts $J(\Sigma_a, R, 0, \infty) < \gamma$. The proof of (b) ⇒ (a) of Theorem 2.3 can be readily completed by reversing the steps in the above argument to establish that the existence of P with the requisite properties is sufficient for $J(\Sigma_a, R, 0, \infty) < \gamma$.

Next we provide a brief sketch of the proof of equivalence of parts (a) and (b) of Theorem 2.2. Suppose $J(\Sigma_a, R, S, T) < \gamma$. It follows that the cost in the optimal control problem

$$\inf_{w \in L_2[0,t]} \left\{ \|w\|_T^2 - \frac{1}{\gamma^2}\left(\|Cx\|_T^2 + x'(T)Sx(T)\right) \right\}$$

subject to the system Σ_a with $x_0 = 0$ is nonnegative. The existence of $P(t)$ satisfying the Riccati differential equation of Theorem 2.2 now follows from standard arguments in classical linear-quadratic control theory (see, e.g., Brockett [3], Limebeer et al. [10]). The rest of the proof is analogous to that of Theorem 2.3.

Equivalence of (a) and (c). We next prove the existence of $Q(t)$ with the desired properties if $J(\Sigma_a, R, S, T) < \gamma$ (respectively, $J(\Sigma_a, R, 0, \infty) < \gamma$). Since H_∞ analysis Riccati equations are seldom written in this form, our proof of this part will be in much greater detail than the one involving $P(t)$ or P. The following arguments also appear implicitly in [12] and [13].

The following Hamiltonian system plays an important role in establishing the existence of $Q(t)$:

$$(10)\begin{pmatrix} \dot{p} \\ \dot{x} \end{pmatrix} = \begin{pmatrix} -A' & -\frac{1}{\gamma^2}C'C \\ BB' & A \end{pmatrix}\begin{pmatrix} p \\ x \end{pmatrix}, \quad \begin{pmatrix} p(0) \\ x(0) \end{pmatrix} = \begin{pmatrix} Rx_0 \\ x_0 \end{pmatrix}, \quad x_0 \in \mathbf{R}^n.$$

Let the transition matrix of this system be given by

$$(11) \qquad \begin{pmatrix} p(t) \\ x(t) \end{pmatrix} = \begin{pmatrix} \Phi_{11}(t,0) & \Phi_{12}(t,0) \\ \Phi_{21}(t,0) & \Phi_{22}(t,0) \end{pmatrix}\begin{pmatrix} p(0) \\ x(0) \end{pmatrix}.$$

In the following, for any $\tau \in [0,T]$ (for the infinite horizon case $\tau \in [0,\infty)$), we will use π_τ to denote the projection operator defined as $(\pi_\tau f)(t) = f(t)$ when $t \le \tau$ and $(\pi_\tau f)(t) = 0$ for $t > \tau$. Suboptimality of J implies that for some $\delta \neq 0$, $J^2(\Sigma_a, R, S, T) < \gamma^2(1-\delta^2)$ (respectively, $J^2(\Sigma_a, R, 0, \infty) < \gamma^2(1-\delta^2)$). Now for any $\tau \in (0,T]$ (respectively, $\tau \in (0,\infty)$), this implies that

$$(12) \qquad x_0'Rx_0 + \|\pi_\tau w\|^2 - \frac{1}{\gamma^2}\|\pi_\tau(Cx)\|^2 \ge \delta^2(x_0'Rx_0 + \|\pi_\tau w\|^2)$$

for all $x(0) \in \mathbf{R}^n$ and $w \in L_2[0,\tau]$ for system (4). From observation (12) we will now show that $[\Phi_{11}(t,0)R + \Phi_{12}(t,0)]$ and $[\Phi_{21}(t,0)R + \Phi_{22}(t,0)]$ are both nonsingular for all $t \in [0,T]$ (respectively, $t \in [0,\infty)$). First, suppose, contrary to what we want to prove, that for some $\tau \in (0,T]$ (respectively, $\tau \in (0,\infty)$), $[\Phi_{11}(\tau,0)R + \Phi_{12}(\tau,0)]$ is singular. Then there exists $x_0 \neq 0$ such that $[\Phi_{11}(\tau,0)R + \Phi_{12}(\tau,0)]x_0 = 0$. Thus for system (10) with $x(0) = x_0$ and $p(0) = Rx_0$, $p(\tau) = 0$. Choosing $w(t) = B'p(t)$ for $t \in [0,\tau]$, systems (4) and (10) have the same trajectory for x. Differentiating the product $x'(t)p(t)$ along the trajectory of (10), we obtain

$$(13) \qquad \frac{d(x'p)}{dt} = w'w - \frac{1}{\gamma^2}x'C'Cx.$$

Integrating (13) from 0 to τ and noting that $p(\tau) = 0$, we get

$$x_0'Rx_0 + \|\pi_\tau w\|^2 - \frac{1}{\gamma^2}\|\pi_\tau(Cx)\|^2 = 0,$$

which is a clear contradiction to (12) since $x(0) = x_0 \neq 0$. Nonsingularity of $[\Phi_{21}(t,0)R + \Phi_{22}(t,0)]$ is shown similarly. (If $[\Phi_{21}(\tau,0)R + \Phi_{22}(\tau,0)]$ is singular for some τ, then with an appropriately chosen $x(0) \neq 0$, $x(\tau)$ is zero and we arrive at a similar contradiction.)

Next, setting $Q(t) := [\Phi_{21}(t,0)R + \Phi_{22}(t,0)][\Phi_{11}(t,0)R + \Phi_{12}(t,0)]^{-1}$, straightforward differentiation shows that $Q(t)$ satisfies the desired Riccati differential equation. Note that $Q(t)$ is nonsingular for all $t > 0$ since $[\Phi_{21}(t,0)R + \Phi_{22}(t,0)]$ is also nonsingular. Since $Q(t)$ is nonsingular for all $t \ge 0$, $Q(0) > 0$, and $Q(t)$ is a continuous function of t, it now follows that $Q(t) > 0$ for all $t > 0$.

For the following proposition, we will need the Lyapunov differential equation

$$\dot{\mathcal{C}}(t) = A\mathcal{C} + \mathcal{C}A' + BB', \qquad \mathcal{C}(0) = R^{-1}. \tag{14}$$

Since A is a constant stable matrix, $\mathcal{C}(t)$ is bounded. Let $\rho \geq \sup_{t>0} \|\mathcal{C}(t)\|$, where the norm of a matrix is defined as the largest singular value.

Before proceeding with the remainder of the proof, we state a straightforward result from quadratic optimization theory which will prove very useful in some remaining parts of the proof. (The following result can be easily proven by standard completion of squares.)

PROPOSITION 1.4. *Consider the system Σ_a given by* (4) (*in the infinite horizon case A is also stable*), *and let $R > 0$ be a given positive-definite matrix. Then for any $\tau \in [0, T]$* (*respectively, $\tau \in [0, \infty)$ for the infinite horizon case*)

$$\inf_{x(0)\in\mathbf{R}^n, w\in\mathcal{L}_2[0,\tau]} \left(\|\pi_\tau w\|^2 + x'(0)Rx(0) - \frac{1}{\gamma^2}\|\pi_\tau z\|^2 \, : \, x(\tau) = x_\tau \right) = x_\tau' Q^{-1}(\tau) x_\tau, \tag{15}$$

$$\inf_{x(0)\in\mathbf{R}^n, w\in\mathcal{L}_2[0,\tau]} (\|\pi_\tau w\|^2 + x'(0)Rx(0) \, : \, x(\tau) = x_\tau) = x_\tau' \mathcal{C}^{-1}(\tau) x_\tau \geq \frac{1}{\rho}\|x_\tau\|^2, \tag{16}$$

where ρ is defined as above.

If $J(\Sigma_a, R, S, T) < \gamma$, then,

$$\inf_{x(0)\in\mathbf{R}^n, w\in\mathcal{L}_2[0,T]} \left(\|\pi_T w\|^2 + x'(0)Rx(0) - \frac{1}{\gamma^2}\|\pi_T z\|^2 \, : \, x(T) = x_T \right) - \frac{1}{\gamma^2} x_T' S x_T > 0$$

for all $x_T \in \mathbf{R}^n$. Invoking the above proposition, it now follows that $\gamma^2 Q^{-1}(T) > S$.

For part (c) of the infinite horizon case (Theorem 2.3), it still remains to be shown that $Q(t)$ is bounded, stabilizing, and asymptotically converges to the stabilizing solution of the corresponding algebraic Riccati equation. We first show that there exists a positive number $\nu < \infty$ such that $Q(t) \leq \nu I$ for all $t > 0$. Based on Proposition 2.1, now consider the following series of inequalities:

$$\begin{aligned} x_\tau' Q^{-1}(\tau) x_\tau &= \inf_{x(0)\in\mathbf{R}^n, w\in\mathcal{L}_2[0,\tau]} \left(\|\pi_\tau w\|^2 + x'(0)Rx(0) - \frac{1}{\gamma^2}\|\pi_\tau z\|^2 : x(\tau) = x_\tau \right) \\ &\geq \delta^2 \inf_{x(0)\in\mathbf{R}^n, w\in\mathcal{L}_2[0,\tau]} (\|\pi_\tau w\|^2 + x'(0)Rx(0) : x(\tau) = x_\tau) \quad \text{from (12)} \\ &= \delta^2 x_\tau' \mathcal{C}^{-1}(\tau) x_\tau \\ &\geq \frac{\delta^2}{\rho} x_\tau' x_\tau. \end{aligned} \tag{17}$$

The above bound, which is independent of τ and true for all $x_\tau \in \mathbf{R}^n$, shows that $Q^{-1}(\tau)$ is bounded below for all $\tau > 0$ or that $Q(\tau)$ is bounded.

Now we show that the time-varying system generated by $A + (1/\gamma^2)Q(t)C'C$ is exponentially stable. The Hamiltonian system associated with the optimization problem

$$\inf_{x(0)\in\mathbf{R}^n, w\in\mathcal{L}_2[0,\tau]} \left(\|\pi_\tau w\|^2 + x'(0)Rx(0) - \frac{1}{\gamma^2}\|\pi_\tau z\|^2 \, : \, x(\tau) = x_\tau \right),$$

which has a solution for all $\tau > 0$ and $x_\tau \in \mathbf{R}^n$ when $J(\Sigma_a, R, 0, \infty) < \gamma$, can be written as

$$\begin{pmatrix} \dot{p} \\ \dot{x} \end{pmatrix} = \begin{pmatrix} -A' & -\frac{1}{\gamma^2}C'C \\ BB' & A \end{pmatrix} \begin{pmatrix} p \\ x \end{pmatrix}, \qquad \begin{pmatrix} p(0) \\ x(\tau) \end{pmatrix} = \begin{pmatrix} Rx(0) \\ x_\tau \end{pmatrix}. \tag{18}$$

The optimal w then for the above optimization problem is $w(t) = B'p(t)$ and it can be easily verified that $x(t) = Q(t)p(t)$ for all $t \in [0, \tau]$. Integrating $d(x'p)/dt$ for the system (18), we get the optimal cost for the optimization problem and similar arguments as above yield

$$\begin{aligned} x'_\tau p(\tau) &= \inf_{x(0)\in\mathbf{R}^n, w\in\mathcal{L}_2[0,\tau]} \left(\|\pi_\tau w\|^2 + x'(0)Rx(0) - \frac{1}{\gamma^2}\|\pi_\tau z\|^2 : x(\tau) = x_\tau \right) \\ &\geq \delta^2(\|\pi_\tau B'p\|^2 + x'(0)Rx(0)) \quad \text{for system (18) (from (12))} \\ &> \delta^2 \frac{1}{\gamma^2}\|\pi_\tau Cx\|^2 \quad \text{for system (18), since } J(\Sigma_a, R, 0, \infty) < \gamma. \end{aligned} \tag{19}$$

Since $x(t) = Q(t)p(t)$, for system (18), the dynamics of p can also be written as

$$\dot{p}(t) = \left[-A' - \frac{1}{\gamma^2}C'(t)C(t)Q(t) \right] p(t). \tag{20}$$

Equations (19) and (20) imply that for all $\tau > 0$ and any $p(\tau) \in \mathbf{R}^n$,

$$x'_\tau p(\tau) = p'(\tau)Q(\tau)p(\tau) > \delta^2 \frac{1}{\gamma^2}\|\pi_\tau CQp\|^2$$

for system (20). Also note that system (20), with the output CQp is detectable (to be thought of here as a system moving backward in time) since A is assumed to be asymptotically stable. Thus the above equation implies that a detectable output for system (20) is *uniformly* bounded with respect to terminal time τ and the terminal condition $p(\tau)$. From Brockett [3] it follows that system (20) is exponentially stable moving backward. Note that the forward-moving system for which we want to establish exponential stability is nothing but the adjoint of system (20).

Next we show that $\lim_{t\to\infty} Q(t)$ exists and equals Q_∞, where Q_∞ is the solution of the algebraic Riccati equation with the properties described in part (c) of Theorem 2.3. Since the system is linear time-invariant, differentiating (8) we obtain

$$\ddot{Q}(t) = \left[A + \frac{1}{\gamma^2}Q(t)C'C \right] \dot{Q}(t) + \dot{Q}(t) \left[A + \frac{1}{\gamma^2}Q(t)C'C \right]'.$$

Since we have already established exponential stability of $A + (1/\gamma^2)Q(t)C'C$, it follows that $\dot{Q}(t)$ approaches 0 exponentially as t goes to ∞. Therefore, $\lim_{t\to\infty} Q(t)$ exists. Let $\lim_{t\to\infty} Q(t) = \bar{Q}$. Then $\bar{Q}$ satisfies the algebraic Riccati equation given in part (c) of Theorem 2.3. By using standard results on the stability of linear time-varying systems that are asymptotically time invariant, it follows that $A + (1/\gamma^2)\bar{Q}C'C$ is asymptotically stable. Since the stabilizing solution of the algebraic Riccati equation is unique, we conclude that $\bar{Q} = Q_\infty$.

It now remains to establish that (c) $\Rightarrow$ (a). We first consider Theorem 2.2. It is easily seen that

$$\frac{d(x'Q^{-1}x)}{dt} = w'w - \frac{1}{\gamma^2}x'C'Cx - [w - B'Q^{-1}x]'[w - B'Q^{-1}x]. \tag{21}$$

Integrating the above from 0 to T, we obtain

$$\begin{aligned} &x'(0)Rx(0) + \|w\|_T^2 - \frac{1}{\gamma^2}[\|z\|_T^2 + x'(T)Sx(T)] \\ (22) \qquad &= x'(T)\left[Q^{-1}(T) - \frac{1}{\gamma^2}S\right]x(T) + \|w - B'Q^{-1}x\|_T^2 \geq 0. \end{aligned}$$

To show that $J(\Sigma_a, R, S, T) < \gamma$, we still need to show that there exists a $\delta > 0$ such that the right-hand side of (22) is greater than or equal to $\delta^2(x_0'Rx_0 + \|w\|_T^2)$. Towards this goal, consider the mapping

$$\Gamma : \mathbf{R}^n \oplus L_2[0,T] \to \mathbf{R}^n \oplus L_2[0,T] : \ (x(0), w) \to (x(T), r),$$

where $r := w - B'Q^{-1}x$. On the domain of Γ, define the inner product:

$$\langle (x_1, w_1), (x_2, w_2) \rangle := x_1'Rx_2 + \langle w_1, w_2 \rangle_{L_2[0,T]},$$

where $\langle \cdot, \cdot \rangle_{L_2[0,T]}$ is the usual inner product in $L_2[0,T]$. On the co-domain of Γ, define the inner product:

$$\langle (x_1, r_1), (x_2, r_2) \rangle := x_1'\left[Q^{-1}(T) - \frac{1}{\gamma^2}S\right]x_2 + \langle w_1, w_2 \rangle_{L_2[0,T]}.$$

It can be easily verified that Γ is an invertible operator. Indeed, let $(x_T, r) \in \mathbf{R}^n \oplus L_2[0,T]$ be given. Then the following system

$$(23) \qquad \begin{aligned} &\dot{x} = (A + BB'Q^{-1})x + Br, \qquad x(T) = x_T, \\ &x(0) = x(0), \qquad w = r + B'Q^{-1}x, \end{aligned}$$

is the inverse of Γ. We will denote the inverse of Γ given by the above equations as Γ^{-1}. Since the problem is over finite horizon, Γ^{-1} has a bounded norm. From (22) we thus obtain

$$\begin{aligned} x'(0)Rx(0) + \|w\|_T^2 - \frac{1}{\gamma^2}[\|z\|_T^2 + x'(T)Sx(T)] &= x'(T)\left[Q^{-1}(T) - \frac{1}{\gamma^2}S\right]x(T) + \|r\|_T^2 \\ &\geq \frac{1}{\|\Gamma^{-1}\|^2}(x'(0)Rx(0) + \|w\|_T^2). \end{aligned}$$

It now follows that $J(\Sigma_a, R, S, T) < \gamma$.

Next we show that if part (c) of Theorem 2.3 is satisfied then $J(\Sigma_a, R, 0, \infty) < \gamma$. If the system Σ_a given in (4) is time invariant and stable, and the inner product on the domain of Σ_a ($\mathbf{R}^n \oplus L_2[0,\infty]$) is the same as in the domain of Γ defined above, then it is easily seen that Σ_a^*, the adjoint of Σ_a is given by

$$\Sigma_a^* : L_2[0,\infty) \to \mathbf{R}^n \oplus L_2[0,\infty),$$

$$(24) \qquad \begin{aligned} &\frac{dx_a}{dt} = -A'x_a - C'w_a, \qquad \lim_{t\to\infty} x_a(t) = 0, \\ &z_a = B'x_a, \\ &(\Sigma_a^* w_a)(t) = \begin{bmatrix} R^{-1}x_a(0) \\ z_a(t) \end{bmatrix}. \end{aligned}$$

(Since A is a Hurwitz matrix, it is not difficult to see that given any w_a in $L_2[0,\infty)$, there is a unique x_a in $L_2[0,\infty)$ that satisfies the state equation in (24).) To show $J(\Sigma_a, R, 0, \infty) < \gamma$ is equivalent to showing that $\|\Sigma_a\| = \|\Sigma_a^*\| < \gamma$. For showing $\|\Sigma_a^*\| < \gamma$, from (24) it follows that we must show

$$\sup_{0 \neq w_a \in \mathcal{L}_2[0,\infty)} \frac{x_a'(0)R^{-1}x_a(0) + \|z_a\|^2}{\|w_a\|^2} < \gamma^2.$$

We easily show that for the adjoint system given by (24),

$$\frac{d(x_a'Q(t)x_a)}{dt} = -\gamma^2 w_a'w_a + z_a'z_a + \left[\gamma w_a - \frac{1}{\gamma}CQ(t)x_a\right]' \left[\gamma w_a - \frac{1}{\gamma}CQ(t)x_a\right].$$

Integrating above from 0 to ∞, we obtain

$$(25) \quad \gamma^2\|w_a\|^2 = \|z_a\|^2 + x_a'(0)R^{-1}x_a(0) + \|v\|^2 \quad \text{where } v := \gamma w_a - \frac{1}{\gamma}CQx_a.$$

The adjoint system (24) can be written in terms of v as

$$\frac{dx_a}{dt} = -\left[A + \frac{1}{\gamma^2}Q(t)C'C\right]' x_a - \frac{1}{\gamma}C'v, \qquad \lim_{t\to\infty} x_a(t) = 0.$$

From the stabilizing property of $A + (1/\gamma^2)Q(t)C'C$, it follows that the above system is exponentially stable moving backward in time. Thus the map from v to w_a is continuous, and hence is bounded, and because of the boundary conditions, $v = 0 \Leftrightarrow w_a = 0$. Thus there exists an $\epsilon > 0$, such that $\|v\|^2 \geq \epsilon\|w_a\|^2$ for all v. This observation together with (25) shows that $\|\Sigma_a^*\| < \gamma$ or, equivalently, $J(\Sigma_a, R, 0, \infty) < \gamma$. □

To make our exposition more lucid, we will make certain simplifying assumptions in this paper. These assumptions are the same as in Doyle et al. [5]. We will assume the following:

(A1) $D_{11} = D_{22} = 0$.

(A2) $D_{12}'[\, C_1 \;\; D_{12}\,] = [\, 0 \;\; I\,]$.

(A3) $D_{21}[\, B_1' \;\; D_{21}'\,] = [\, 0 \;\; I\,]$.

For the infinite horizon case, we will also assume that

(A4) (A, B_1, C_1) is stabilizable and detectable.

(A5) (A, B_2, C_2) is stabilizable and detectable.

Assumption (A1) implies that there is no direct transmission from the exogenous input to regulated output and from the control input to the measured output. The latter assumption ensures that any proper controller leads to a well-posed feedback system. Assumption (A2) is quite common in the LQG literature and is tantamount to assuming that there is no *cross term* in the expression for $\|z\|^2$ and that the penalty on the control input u is normalized, i.e.,

$$\|z\|^2 = \|C_1x\|^2 + \|u\|^2.$$

Assumption (A3) is the dual of assumption (A2) and is analogous to the standard assumption in the Kalman filtering problem that the process noise and the measurement noise are uncorrelated and that the measurement noise is nonsingular and normalized. Assumption (A4) is a technical assumption and is a sufficient condition for certain systems to not have invariant zeros on the imaginary axis. Assumption (A5) is necessary and sufficient to guarantee the existence of an internally stabilizing controller for the system Σ. Assumption (A5) is not required in the finite horizon problems. The reader is referred to Glover and Doyle [7], [8], Safonov, Limebeer, and Chiang [14], and Zhou and Khargonekar [19] on techniques for removing assumptions (A1)–(A4).

2. Main results.

In this section, we state the principal contributions of this paper. Each result presents a necessary and sufficient condition for the solvability of a problem. These necessary and sufficient conditions are stated in terms of existence and properties of solutions to certain (algebraic and differential) Riccati equations. In the event the necessary conditons are met, a state-space formula for a particular (central) solution to the problem is given.

The first result treats the finite horizon state-feedback/full-information problem.

THEOREM 2.1. *Consider the system Σ as in* (1) *with $y = (x' \;\; w')'$. Let R, S be given symmetric matrices such that S is positive semidefinite and R is positive definite.*

(a) *There exists an admissible controller K such that*

$$J(\Sigma_{cl}, R, S, T) < \gamma$$

if and only if there exists a (*unique*) *symmetric matrix function $P(t), t \in [0, T]$ such that*

$$(26) \quad -\dot{P}(t) = A'P(t) + P(t)A + P(t)\left[\frac{1}{\gamma^2}B_1B_1' - B_2B_2'\right]P(t) + C_1'C_1, P(T) = S$$

and $P(0) < \gamma^2 R$.

(b) *In this case, the control law*

$$u(t) = -B_2'P(t)x(t) \tag{27}$$

achieves $J(\Sigma_{cl}, R, S, T) < \gamma$.

The next result addresses the state-feedback/full-information problem in the infinite horizon case.

THEOREM 2.2. *Consider the system Σ as in* (1) *with $y = (x' \;\; w')'$. Let R be a given positive-definite matrix.*

(a) *There exists an admissible controller K such that*

$$J(\Sigma_{cl}, R, 0, \infty) < \gamma$$

if and only if there exists a unique symmetric matrix P such that

$$(28) \quad \begin{array}{ll} \text{(i)} & A'P + PA + P\left(\frac{1}{\gamma^2}B_1B_1' - B_2B_2'\right)P + C_1'C_1 = 0, \\ \text{(ii)} & A + \left(1/\gamma^2)B_1B_1' - B_2B_2'\right)P \quad \textit{is asymptotically stable, and} \\ \text{(iii)} & 0 \leq P < \gamma^2 R. \end{array}$$

(b) *In this case, the control law*

$$u(t) = -B_2'Px(t) \tag{29}$$

achieves $J(\Sigma_{cl}, R, 0, \infty) < \gamma$.

In Theorems 2.1 and 2.2, the necessary conditions are established under the assumption that the controller has access to both x and w. Thus, the necessary conditons would still apply even if the controller has access only to the state x. On the other hand, if the necessary conditons for the existence of controllers are met, one control law that achieves specified performance bounds uses only the knowledge of the state x.

It is interesting to note that as γ approaches ∞, the solution given above approaches the standard solution to the finite and infinite horizon linear-quadratic regulator problem. For example, the Riccati differential equation in Theorem 2.1 approaches the corresponding Riccati equation of the linear-quadratic regulator problem with a terminal cost. Also, the necessary condition $P(0) < \gamma^2 R$ is trivially satisfied as γ approaches ∞. This is intuitively appealing since as γ approaches ∞, the performance requirement $J(\Sigma_{cl}, R, S, T) < \gamma$ becomes trivial, and the solution approaches the solution to the optimal linear-quadratic regulator problem. Analogous comments apply to the infinite-horizon case and also to the results to follow. Similar behavior has been noted previously in Doyle et al. [5] for central solutions to the standard H_∞ control problems.

Another interesting feature of Theorems 2.1 and 2.2 is the decoupling in the necessary conditions. The matrix R plays no role in the solutions of the (algebraic and differential) Riccati equations. In fact, if the required solutions exist, they are unique. Once these solutions have been found, we need to check whether $P(0) < \gamma^2 R$ (respectively, $0 \leq P < \gamma^2 R$) hold. It is interesting to note that the control law does not depend on R. As will be discussed below, the situation is quite different in output-feedback problems. There the control law depends on R explicitly. An intuitive explanation of this feature will be given following the statement of Theorem 3.4.

We next consider the output-feedback problem on a finite horizon.

THEOREM 2.3. *Consider the system Σ as in* (1). *Let R, S be given symmetric matrices such that S is positive semidefinite and R is positive definite.*

(a) *There exists an admissible output feedback controller K such that with the control law $u = Ky$,*

$$J(\Sigma_{cl}, R, S, T) < \gamma$$

if and only if the following three conditions hold:

(i) *There exists a symmetric matrix function $P(t)$ such that*

$$(30)\quad \begin{gathered} -\dot{P}(t) = A'P(t) + P(t)A + P(t)\left[\frac{1}{\gamma^2}B_1B_1' - B_2B_2'\right]P(t) + C_1'C_1, \\ P(T) = S \end{gathered}$$

and $P(0) < \gamma^2 R$.

(ii) *There exists a symmetric matrix function $Q(t) > 0$ for all $t \in [0, T]$ such that*

$$(31)\dot{Q}(t) = AQ(t) + Q(t)A' - Q(t)\left[C_2'C_2 - \frac{1}{\gamma^2}C_1C_1'\right]Q(t) + B_1B_1',$$

with $Q(0) = R^{-1}$.

(iii) *$\rho(1/\gamma^2 P(t)Q(t)) < 1$ for all $t \in [0, T]$.*

(b) *If the above conditions are met, then one controller that achieves*

$$J(\Sigma_{cl}, R, S, T) < \gamma$$

is given by

$$\begin{aligned} \frac{d\hat{x}(t)}{dt} &= \left[A + \frac{1}{\gamma^2} B_1 B_1' P\right] \hat{x}(t) + \left[I - \frac{1}{\gamma^2} Q(t) P\right]^{-1} \\ &\quad \cdot Q(t) C_2' [y(t) - C_2 \hat{x}(t)] + B_2 u(t), \\ \hat{x}(0) &= 0, \\ u(t) &= -B_2' P \hat{x}(t), \end{aligned} \tag{32}$$

(In the above theorem $\rho(\cdot)$ denotes the spectral radius.)

The next result gives a solution for the output-feedback problem in the infinite horizon case for linear time-invariant systems.

THEOREM 2.4. *Consider the system Σ as in* (1). *Let R be a given positive-definite matrix.*

(a) *There exists an admissible output-feedback controller K such that with the control law $u = Ky$,*

$$J(\Sigma_{cl}, R, 0, \infty) < \gamma$$

if and only if the following three conditions hold:

(i) *There exists a* (*unique*) *symmetric matrix P such that*

$$A'P + PA + P\left(\frac{1}{\gamma^2} B_1 B_1' - B_2 B_2'\right) P + C_1' C_1 = 0, \tag{33}$$

with $A + ((1/\gamma^2) B_1 B_1' - B_2 B_2') P$ being asymptotically stable and $\gamma^2 R > P \geq 0$.

(ii) *There exists a symmetric bounded matrix function $Q(t) > 0$ for all $t \geq 0$ such that*

$$\dot{Q}(t) = AQ(t) + Q(t)A' + [\frac{1}{\gamma^2} C_1 C_1' - Q(t)[C_2' C_2] Q(t) + B_1 B_1', Q(0) = R^{-1}. \tag{34}$$

and the unforced linear time-varying system

$$\dot{p}(t) = \left[A - Q(t)\left(C_2' C_2 - \frac{1}{\gamma^2} C_1' C_1\right)\right] p(t) \tag{35}$$

is exponentially stable.

(iii) *The function $(1 - \rho((1/\gamma^2) Q(t) P))^{-1} > 0$ for all $t \geq 0$ and is bounded.*

(b) *Moreover, if $Q(t)$ with above properties exists for all $t \geq 0$, then $\lim_{t \to \infty} Q(t)$ exists and equals Q_∞, where Q_∞ is the unique symmetric matrix with the following properties:*

$$A Q_\infty + Q_\infty A' - Q_\infty \left(C_2' C_2 - \frac{1}{\gamma^2} C_1' C_1\right) Q_\infty + B_1 B_1' = 0, \tag{36}$$

$A - Q_\infty \left[C_2' C_2 - (1/\gamma^2) C_1' C_1\right]$ is asymptotically stable, and $Q_\infty \geq 0$.

(c) *If the conditions above are met, then one controller that achieves $J(\Sigma_{cl}, R, 0, \infty) < \gamma$ is given by*

$$\begin{aligned} \frac{d\hat{x}(t)}{dt} &= \left[A + \frac{1}{\gamma^2}B_1B_1'P\right]\hat{x}(t) + \left[I - \frac{1}{\gamma^2}Q(t)P\right]^{-1} \\ &\quad \cdot Q(t)C_2'[y(t) - C_2\hat{x}(t)] + B_2u(t), \\ \hat{x}(0) &= 0 \\ u(t) &= -B_2'P\hat{x}(t), \end{aligned} \tag{37}$$

It should be noted that as in Doyle et al. [5], the necessary and sufficient conditions for the existence of controllers in the output-feedback problems are two decoupled Riccati equations and a "spectral radius" condition.

From the formulae for the controllers, we note that in contrast to the state-feedback problems, the initial condition weighting matrix R plays an important role in the controller formulae in the output-feedback problems. An intuitive explanation for this phenomenon is as follows. In the state-feedback problem, since the entire state is available for feedback, there is no uncertainty in the initial state as far as the controller is concerned. The only issue in the state-feedback case is whether the desired performance bound is achievable. This is verified by checking the inequalities $P(0) < \gamma^2 R$ in the finite-horizon case and $0 < P < \gamma^2 R$ in the infinite-horizon case. On the other hand, in the output-feedback case, the controller does not have complete knowledge of the initial state. Consequently, the controller gains depend on the relative weighting between the uncertainty in the initial state vis-à-vis that in the exogenous input w to satisfy the desired performance bound for all w and $x(0)$. Also, the controller gains change with time reflecting the relative importance of the information contained in measurements as compared to the prior information on the initial state. Finally, we would like to note that in the infinite time horizon case, as $t \to \infty$, the controller approaches the central controller for an associated standard H_∞ problem. This is intuitively appealing since assuming that the controller is internally stabilizing, as $t \to \infty$, the effect of nonzero initial states should disappear and the controller should be required only to attenuate the effect of exogenous signals w on z. This is exactly how the controller given in Theorem 2.4 behaves.

For sufficiently large R, the steady-state central controller itself has sufficient robustness to account for uncertainty due to unknown initial condition and we need not necessarily require a linear time-varying controller to achieve the desired performance. This is formally stated in the following result.

COROLLARY 2.5. *Let the conditions of Theorem* 3.4 *be satisfied and* R *be such that* $R^{-1} < Q_\infty$. *Then the linear time-invariant controller of the form given in Theorem* 3.4(c) *with* Q_∞ *replacing* $Q(t)$ *achieves* $J(\Sigma_{cl}, R, 0, \infty) < \gamma$.

3. Proofs.

Without loss of generality, we will assume that $\gamma = 1$.

Proof of Theorem 3.2. Suppose that both x, w are available as measurements to the controller and suppose that there exists a controller such that $J(\Sigma_{cl}, R, 0, \infty) < 1$. Then it follows from Lemma 2.1 that for the system Σ with $x_0 = 0$, we have

$$\sup_w \inf_u \frac{\|z\|}{\|w\|} < 1.$$

From Doyle et al. [5], it follows that there exists $P \geq 0$ such that P satisfies (28) and $(A + B_1B_1'P - B_2B_2'P)$ is asymptotically stable. It remains to be shown that $P < R$. We will prove this by contradicition.

Suppose that there exists $x_0 \neq 0$ such that $x_0'(P - R)x_0 \geq 0$. Now in the system Σ, set $x(0) = x_0$ and $w(t) = B_1'Pe^{(A+(B_1B_1'-B_2B_2')P)t}x_0$. Using the fact that

$A + (B_1B_1' - B_2B_2')P$ is asymptotically stable, it follows that w belongs to $L_2[0, \infty)$. For this particular w, we claim that

$$\sup_u \{ \ \|w(t)\|^2 - \|z(t)\|^2 \ \} = -x_0'Px_0.$$

Indeed, the (unique) optimal input u for the above optimal control problem and the corresponding state trajectory are

$$u(t) = \hat{u}(t) = -B_2'Pe^{(A+(B_1B_1'-B_2B_2')P)t}x_0,$$

$$\hat{x}(t) = e^{(A+(B_1B_1'-B_2B_2')P)(t-\tau)}x_0.$$

This can be shown, for example, by noting that for any $w \in L_2[0, \infty)$, the optimal u denoted by $\hat{u}$ is obtained from the following two-point boundary value problem

$$\begin{pmatrix} \dot{p} \\ \dot{x} \end{pmatrix} = \begin{pmatrix} -A' & C_1'C_1 \\ B_2B_2' & A \end{pmatrix} \begin{pmatrix} p \\ x \end{pmatrix} + \begin{pmatrix} 0 \\ B_1w \end{pmatrix}, \quad x(0) = x_0, \quad \lim_{t\to\infty} p(t) = 0,$$

$$\hat{u}(t) = B_2'p(t).$$

The above claim now follows by directly verifying that $p(t) = -Pe^{(A+(B_1B_1'-B_2B_2')P)t}x_0$. (Similar optimization plays an important role in Tadmor [15].)

Using this observation and the fact that $J(\Sigma_{cl}, R, 0, \infty) < 1$, it follows that

$$0 < x_0'Rx_0 + \|w\|^2 - \|z\|^2 \leq x_0'Rx_0 + \sup_u\{\|w\|^2 - \|z\|^2\} = x_0'(R-P)x_0 \leq 0,$$

which is a contradiction.

If there exists a $P \geq 0$ such that P satisfies (28) and $(A + B_1B_1'P - B_2B_2'P)$ is asymptotically stable, then using detectability of (A, C_1), standard Lyapunov techniques and a completion of squares argument together with $P < R$ can be employed to show that the control law (27) is internally stabilizing and achieves the desired performance bound. □

Proof of Theorem 3.1. Suppose that both x, w are available to the controller and suppose that there exists a controller such that $J(\Sigma_{cl}, R, S, T) < 1$. Now consider the system Σ with $x_0 = 0$. Then it follows that

$$\sup_w \inf_u \frac{\|z\|_T^2 + x(T)'Sx(T)}{\|w\|_T^2} < 1.$$

(The supremum and the infimum above are conducted over signals $w, u \in L_2[0, T]$.) Using the approach of Limebeer et al. [10] (which involves conjugate point analysis for a certain two-point boundary value problem associated with the quadratic game), it can be shown that there exists a $P(t)$ that satisfies (26). The details are omitted for the sake of brevity.

To complete the necessity it remains to be shown that $P(0) < R$. Suppose to the contrary that there exists $x_0 \neq 0$ such that $x_0'(R - P(0))x_0 \leq 0$. By differentiating $x'Px$ along the trajectories of Σ, and completing squares we obtain

$$(38) \quad \frac{d(x'Px)}{dt} = [u + B_2'Px]'[u + B_2'Px] - [w - B_1'Px]'[w - B_1'Px] + w'w - z'z.$$

With $x(0) = x_0$ and $w(t) = B_1'P(t)x(t)$, integrating the expression above from 0 to T yields

$$0 < x'(0)Rx(0) + \|w\|_T^2 - \|z\|_T^2 - x'(T)Sx(T) = x_0'(R-P)x_0 - \|u + B_2'Px\|_T^2 \leq 0,$$

which is a contradiction.

The proof of sufficiency in Theorem 3.1 follows readily by integrating (38) from 0 to T, and is omitted. □

Proofs of Theorems 3.3, 3.4, *and Corollary* 3.5.

Sufficiency proofs for Theorems 3.3 *and* 3.4. Here we show that if the solutions to the corresponding algebraic and differential Riccati equations of Theorems 3.3 and 3.4 exist and satisfy the required properties, then the controllers given by (32) and (37) achieve the desired performance bound. We consider the finite horizon and infinite horizon cases separately.

Finite horizon case. Integrating (38) from 0 to T, it is easily seen that

$$\begin{aligned} &x'(0)Rx(0) + \|w\|_T^2 - \|z\|_T^2 - x'(T)Sx(T) \\ &= x'(0)(R - P(0))x(0) + \|w - B_1'Px\|_T^2 - \|u + B_2'Px\|_T^2. \end{aligned} \tag{39}$$

We will show that if we use the controller (32) then there exists an $\epsilon > 0$ such that the right-hand side in the above equation is no less than $\epsilon(x'(0)Rx(0) + \|w\|_T^2)$. This will establish our claim that this controller achieves $J(\Sigma_{cl}, R, S, T) < 1$.

Let $Z(t) := [I - Q(t)P(t)]^{-1}Q(t)$. From conditions (a)-(i) and (a)-(ii) of Theorem 3.3, we obtain

$$\begin{aligned} \dot{Z}(t) &= [A + B_1B_1'P(t)]Z(t) + Z(t)[A + B_1B_1'P(t)]' \\ &\quad + Z(t)[P(t)B_2B_2'P(t) - C_2'C_2]Z(t) + B_1B_1' \end{aligned} \tag{40}$$

with

$$Z(0) = [R - P(0)]^{-1}.$$

Let $e := x - \hat{x}$, $r := w - B_1'Px$, $\nu := u + B_2'Px$ and $A_{tmp} := A + B_1B_1'P$. Consider the mapping

$$\Gamma_a : \mathbf{R}^n \oplus L_2[0,T] \to L_2[0,T] : \ (x(0), r) \to \nu.$$

In $\mathbf{R}^n \oplus L_2[0,T]$, (the domain of Γ_a), we define the following inner product:

$$\langle (x_1, r_1), (x_2, r_2) \rangle := x_1'(R - P(0))x_2 + \langle r_1, r_2 \rangle_{L_2[0,T]},$$

where $\langle \cdot, \cdot \rangle_{L_2[0,T]}$ is the usual inner product in $L_2[0,T]$. Fairly straightforward algebra reveals that if we use the controller (32), the system Γ_a can be realized as

$$\dot{e} = (A_{tmp} - ZC_2'C_2)e + (B_1 - ZC_2'D_{21})r, e(0) = x(0), \tag{41}$$

$$\nu = B_2'Pe.$$

It then follows from (40) and Theorem 2.2 that $J(\Gamma_a, R - P(0), 0, T) < 1$. This implies that if we use the controller (32) then there exists an $\epsilon > 0$ such that for all $(x(0), r) \in \mathbf{R}^n \oplus L_2[0,T]$,

$$x'(0)(R-P)x(0) + \|r\|_T^2 - \|\nu\|_T^2 \geq \epsilon(x'(0)(R-P)x(0) + \|r\|_T^2). \tag{42}$$

Next observe simply from the definition of r that the map $(x(0), w) \to (x(0), r)$ is an invertible finite-dimensional linear system and *since we are dealing with finite horizon problems,* this map and its inverse both are bounded. Thus there exists a $M > 0$ such that for all $(x(0), w) \in \mathbf{R}^n \oplus L_2[0, T]$,

$$x'(0)(R-P)x(0) + \|r\|_T^2 \geq M(x'(0)Rx(0) + \|w\|_T^2). \tag{43}$$

Combining the above two observations, we conclude that, using the controller (32),

$$x'(0)(R-P)x(0)+\|r\|_T^2-\|\nu\|_T^2 \geq \epsilon(x'(0)(R-P)x(0)+\|r\|_T^2) \geq \epsilon M(x'(0)Rx(0)+\|w\|_T^2) \tag{44}$$

for all $(x(0), w) \in \mathbf{R}^n \oplus L_2[0, T]$. From (39) It follows that that this controller achieves $J(\Sigma_{cl}, R, S, T) < 1$.

Infinite horizon case. Here we first need to show the internal stability of the closed-loop system. This is actually quite easy. Note that as t approaches ∞, the controller (37) approaches a linear time-invariant controller. Indeed, from condition (b) in the Theorem 3.4, it follows that the controller (37) approaches the so-called central solution in Doyle et al. [5]. It follows from Theorem 3 of Doyle et al. [5] that the asymptotic limit of the controller (37) internally stabilizes the system Σ. Now using standard results on the stability of linear time-varying systems, which are asymptotically time-invariant, it follows that (37) internally exponentially stabilizes Σ.

Showing that the controller (37) achieves $J(\Sigma_{cl}, R, 0, \infty) < 1$ follows the same lines as the finite horizon case. The only place where the finite horizon case arguments need further elaboration to be applicable here is in establishing (43). For establishing (43), it is sufficient to show that the operator from $(x(0), r) \to (x(0), w)$ is bounded. Toward that end we first show that the system Γ_a described by (41) is exponentially stable. Equation (40) can also be written as

$$\begin{aligned}\dot{Z}(t) = [A_{tmp} - ZC_2'C_2]Z(t) - Z(t)[A_{tmp} - ZC_2'C_2]' \\ +Z(t)[P(t)B_2B_2'P(t) + C_2'C_2]Z(t) + B_1B_1',\end{aligned}$$

where $A_{tmp} := A + B_1B_1'P(t)$. Since (A, B_1) is stabilizable, it follows that the pair $([A_{tmp} - ZC_2'C_2], [ZC_2' \;\; B_1])$ is also stabilizable. Existence of a bounded positive semidefinite Z with this stabilizability observation and an extension of the standard lemma of Lyapunov (see, Ravi, Nagpal, and Khargonekar [13]) implies that $A_{tmp} - ZC_2'C_2$ is exponentially stable.

With the controller (37) applied to the plant Σ, the closed-loop system equations can be written as (with $e := x - \hat{x}$)

$$\begin{aligned}\begin{pmatrix} \dot{x} \\ \dot{e} \end{pmatrix} = \begin{pmatrix} A + (B_1B_1' - B_2B_2')P & B_2'P \\ 0 & A_{tmp} - ZC_2'C_2 \end{pmatrix} \begin{pmatrix} x \\ e \end{pmatrix} \\ + \begin{pmatrix} B_1 \\ B_1 - ZC_2'D_{21} \end{pmatrix} r.\end{aligned} \tag{45}$$

Since $A + (B_1B_1' - B_2B_2')P$ and $A_{tmp} - ZC_2'C_2$ are both exponentially stable, the above system is exponentially stable. Thus the operator from $(x(0), r) \to (x(0), w)$ is bounded.

Proof of Corollary 3.5. The above arguments show that the linear time-invariant controller obtained by substituting Q_∞ in (37) for $Q(t)$ achieves $J(\Sigma_{cl}, Q_\infty^{-1}, 0, \infty) < 1$.

Since $R > Q_\infty^{-1}$, from Lemma 2.1 it follows that $J(\Sigma_{cl}, R, 0, \infty) \le J(\Sigma_{cl}, Q_\infty^{-1}, 0, \infty) < 1$.

Necessity of Theorems 3.3 *and* 3.4. Given a controller for which $J(\Sigma_{cl}, R, S, T) < \gamma$, (respectively, $J(\Sigma_{cl}, R, 0, \infty) < \gamma$,) condition (i) in Theorems 3.3 and 3.4 follows from the proof of Theorems 3.1 and 3.2.

As before, for any $\tau \in (0, T]$ (for the infinite horizon case $\tau \in (0, \infty)$), let π_τ denote the projection operator defined as $(\pi_\tau f)(t) = f(t)$ when $t \le \tau$ and $(\pi_\tau f)(t) = 0$ for $t > \tau$.

To show the existence of (31) and (34), we will use the results from Nagpal and Khargonekar [12]. We begin this with a simple but important observation. Suppose that $y(t) = 0$ for $t \in [0, \tau]$. Then since the controller is linear and causal, $u(t)$ must equal 0 for $t \in [0, \tau]$. (This property holds for internally stabilizing nonlinear time-varying causal controllers as well where internal stability is taken to mean finite gain L_2 stability.) A similar observation plays a central role in the work of Nagpal and Khargonekar [12] and this allows us to make use of some of their results in the present context.

Let $D_{21}^{\perp}$ be such that

$$\begin{bmatrix} D_{21} \\ D_{21}^{\perp} \end{bmatrix} \begin{bmatrix} D_{21} \\ D_{21}^{\perp} \end{bmatrix}' = I. \tag{46}$$

Such a $D_{21}^{\perp}$ exists due to assumption (A3). For any $\tau \in (0, T]$ (for the infinite horizon case $\tau \in (0, \infty)$), define

$$\text{(47)}\ W(\tau) := \{w \in L_2[0, T] : w(t) = -D_{21}' C_2 x(t) + D_{21}^{\perp\prime} v(t), t \in [0, \tau], v \in L_2[0, \tau]\}.$$

For any $w \in W(\tau)$, we easily observe that for all $t \in [0, \tau]$, $y(t) = 0$. Therefore, $u(t) = 0$, for all $t \in [0, \tau]$. Consequently, for $w \in W(\tau)$, the equations for (1) in the interval $t \in [0, \tau]$ become

$$\begin{aligned} \frac{dx}{dt} &= Ax + B_1 D_{21}^{\perp\prime} v, \qquad t \in [0, \tau], \\ z &= C_1 x, \\ y &= 0, \\ \|\pi_\tau w\|^2 &= \|\pi_\tau (C_2 x)\|^2 + \|\pi_\tau v\|^2. \end{aligned} \tag{48}$$

Since we are dealing with existence of suboptimal controllers, if there exists a controller such that $J^2(\Sigma_{cl}, R, 0, \infty) < 1$, then there also exists a controller such that $J^2(\Sigma_{cl}, R, 0, \infty) < 1 - \delta^2$ for some $\delta \neq 0$. Now for any $\tau \in (0, T]$ (for the infinite horizon case $\tau \in (0, \infty)$), let $w \in W(\tau)$. Then existence of a controller that achieves $J^2(\Sigma_{cl}, R, 0, \infty) < 1 - \delta^2$ implies that

$$\begin{aligned} & x_0' R x_0 + \|\pi_\tau v\|^2 + \|\pi_\tau (C_2 x)\|^2 - \|\pi_\tau (C_1 x)\|^2 \\ & \ge \delta^2 (x_0' R x_0 + \|\pi_\tau v\|^2 + \|\pi_\tau (C_2 x)\|^2) \end{aligned} \tag{49}$$

for all $x(0) \in \mathbf{R}^n$ and $v \in L_2[0, \tau)$ for system (48). From Nagpal and Khargonekar [12, Thms. 3, 4], the above inequality implies (a)-(ii) for Theorem 3.3 and (a)-(ii) and (b) for Theorem 3.4.

To obtain conditions (a)-(iii) of Theorems 3.3 and 3.4, we use the "separation" idea from Khargonekar [9]. This "separation" idea is roughly as follows : 1) We separate the time interval of the original problem into two subintervals; 2) during the first part, w is chosen so that $y \equiv 0$ (thus during this interval $u \equiv 0$); and 3) in the second interval, one chooses the "worst" w in an appropriate sense (which will soon become clear). Such a choice of w together with the inequality implied by the existence of a controller that achieves $J(\Sigma_{cl}, R, S, T) < 1$ (respectively, $J(\Sigma_{cl}, R, 0, \infty) < 1$ for the infinite horizon case), would lead us to conclude conditions (iii) of the two theorems.

Fix $\tau \in (0, T]$ (respectively, $\tau \in (0, \infty)$) and $x_\tau \in \mathbf{R}^n$. Let $w \in W(\tau)$ such that $v(t) = D_{21}^{\perp}B_1'Q^{-1}(t)x(t)$ for $t \in [0, \tau]$. With this choice of $x(\tau)$ and $v(t)$, it is easily seen by completion of squares that for system (48),

$$\begin{aligned} x_\tau'Q^{-1}(\tau)x_\tau &= x(0)'Rx(0) + \|\pi_\tau v\|^2 + \|\pi_\tau(C_2x)\|^2 - \|\pi_\tau(C_1'x)\|^2 \\ &= x(0)'Rx(0) + \|\pi_\tau w\|^2 - \|\pi_\tau z\|^2. \end{aligned} \tag{50}$$

We now consider the finite and the infinite horizon cases separately.

Finite horizon case. Fix $\tau \in (0, T]$ and $x_\tau \in \mathbf{R}^n$ and let $w \in W(\tau)$ be chosen as in the last paragraph. For $t \in (\tau, T]$, let $w(t) = B_1'P(t)x(t)$. Now integrating (38) from τ to T, we get

$$\begin{aligned} &\|(I - \pi_\tau)w\|^2 - \|(I - \pi_\tau)z\|^2 - x(T)'Sx(T) \\ &= -x_\tau'P(\tau)x_\tau - \|(I - \pi_\tau)(u + B_2'Px)\|^2. \end{aligned} \tag{51}$$

Combining (50) and (51) we obtain

$$\begin{aligned} 0 &\le x_0'Rx_0 + \|w\|_T^2 - \|z\|_T^2 - x(T)'Sx(T) \\ &= \{x(0)'Rx(0) + \|\pi_\tau w\|^2 - \|\pi_\tau z\|^2\} + \{\|(I - \pi_\tau)w\|^2 - \|(I - \pi_\tau)z\|^2 - x(T)'Sx(T)\} \\ &= x_\tau'[Q^{-1}(\tau) - P(\tau)]x_\tau - \|(I - \pi_\tau)(u + B_2'Px)\|^2. \end{aligned}$$

As the first inequality is *strict* for all $0 \ne x_\tau \in \mathbf{R}^n$, (since in this case either $w \ne 0$ or $x(0) \ne 0$), condition (iii) of Theorem 3.3 follows.

Infinite horizon case. Fix $\tau \in (0, \infty)$ and $x_\tau \in \mathbf{R}^n$ and let $w \in W(\tau)$ be chosen as in the finite horizon case. For $t \in (\tau, \infty)$, let $w(t) = B_1'Pe^{(A+(B_1B_1'-B_2B_2')P)(t-\tau)}x(\tau)$. Using the fact that $A + (B_1B_1' - B_2B_2')P$ is asymptotically stable, it follows that $(I - \pi_\tau)w$ belongs to $L_2[0, \infty)$. For this particular w as in the proof of Theorem 3.2, it can be shown that

$$\sup_u\{\|(I - \pi_\tau)w(t)\|^2 - \|(I - \pi_\tau)z(t)\|^2\} = -x'(\tau)P(\tau)x(\tau).$$

Using this observation and (50), we obtain

$$\begin{aligned} x_0'Rx_0 + \|w\|^2 - \|z\|^2 &= \{x(0)'Rx(0) + \|\pi_\tau w\|^2 - \|\pi_\tau z\|^2\} \\ &\quad+\{\|(I - \pi_\tau)w\|^2 - \|(I - \pi_\tau)z\|^2\} \\ &\le x_\tau'Q^{-1}(\tau)x_\tau - x_\tau'Px_\tau. \end{aligned}$$

Since the left-hand side is positive for all $0 \ne x_\tau \in \mathbf{R}^n$ (since that implies $w \ne 0$ or $x(0) \ne 0$), it follows that for all $\tau \in [0, \infty)$, $Q^{-1}(\tau) > P$. Hence $\rho(Q(\tau)P) < 1$ for all $\tau \in [0, \infty)$. Now $\lim_{t \to \infty} Q(\tau) = Q_\infty$. From Theorem 3 of Doyle et al. [5], $\rho(Q_\infty P) < 1$. Thus, $(1 - \rho(Q(\tau)P))^{-1}$ exists for all $\tau \in [0, \infty)$, is continuous, and has a limit as $\tau \to \infty$. Therefore it is bounded on $[0, \infty)$. Hence condition (iii) of the theorem follows. This completes the proof. □

4. Conclusions. In this paper we have formulated and solved an H_∞-like control problem, where, in addition to the exogenous signals in the state and measurement equations, we must also account for uncertainty in the initial condition. This was done by defining a suitable worst-case performance measure. It is hoped that these results may allow us to design control strategies that are robust to both exogenous signals and nonzero initial states.

It was shown by Mustafa and Glover [11] that the central solutions to H_∞ control problems have the additional property of maximizing the entropy of the closed-loop transfer matrix. It is clear that the controllers obtained in the present paper are analogous to the central solutions to the H_∞ control problem. The question arises then as to what is the analogue of entropy maximization property in the present context. This problem does not seem to have immediate answers since entropy is defined in terms of the closed-loop transfer function matrix evaluated along the imaginary axis. This definition has no obvious extension to our context where the controllers are time-varying. We leave a full investigation of this issue for future research.

Finally, we have formulated our problem in a purely deterministic context. A stochastic formulation of this problem should be an interesting undertaking.

Acknowledgment. The authors are grateful to R. Ravi for many helpful discussions regarding the content of this paper.

REFERENCES

[1] B. D. O. Anderson, *An algebraic solution to the spectral factorization problem,* IEEE Trans. Automat. Control, AC-12 (1967), pp. 410–414.

[2] S. Boyd, V. Balakrishnan, and P. Kabamba, *A bisection method for computing H_∞ norm of a transfer matrix and related problems*, Mathematics Control Signals Systems, 2 (1989), pp. 207–219.

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

[4] G. Didinsky and T. Basar, *Design of minimax controllers for linear systems with nonzero initial conditions and under specified information structures*, preprint, Coordinated Science Laboratory, University of Illinois, Urbana, IL,1990.

[5] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, *State-space solutions to standard H_2 and H_∞ and control problems,* IEEE Trans. Automat. Control., 34 (1989), pp. 831–847.

[6] B. A. Francis, *A Course in H_∞ Control Theory,* Lecture Notes in Control and Inform. Sci., Vol. 88, Springer-Verlag, New York, 1988.

[7] K. Glover and J. C. Doyle, *State-space formulae for all stabilizing controllers that satisfy an H_∞ norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[8] ———, *A state-space approach to H_∞ optimal control*, in Three Decades of Mathematical System Theory, H. Nijmeijer and J. M. Schumacher, eds., Springer-Verlag, Berlin, 1989, pp. 179–218.

[9] P. P. Khargonekar, *State-space H_∞ control theory*, in Mathematical System Theory: The Influence of R. E. Kalman, A. C. Antoulas, ed., Springer-Verlag, Berlin, New York, to appear.

[10] D. J. N. Limebeer, B. D. O. Anderson, P. P. Khargonekar, and M. Green, *A game theoretic approach to H_∞ control for time varying systems*, SIAM J. Control Optim., to appear.

[11] D. Mustafa and K. Glover, *Controllers which satisfy an H_∞-norm bound and maximize an entropy integral*, in Proc. 1988 Conference on Decision and Control, pp. 959–964.

[12] K. M. Nagpal and P. P. Khargonekar, *Filtering and smoothing in an H_∞ setting*, IEEE Trans. Automat. Control., 36 (1991), pp. 152–166

[13] R. Ravi, K. M. Nagpal, and P. P. Khargonekar, *H_∞ control for linear time varying systems: a state-space approach*, Control Group Report No. CGR-43, College of Engineering, University of Michigan, Ann Arbor, MI, March 1990; in Proc. 29th IEEE Conference on Decision and Control, December 1990 and SIAM J. Control Optim., this issue, pp. 1394–

1413.

[14] M. G. SAFONOV, D. J. N. LIMEBEER, AND R. Y. CHIANG, *Simplifying the H_∞ theory via loop shifting, matrix pencil, and descriptor concepts*, Internat. J. Control, 50 (1989), pp. 2467–2488.

[15] G. TADMOR , *Worst-case design in the time domain: the maximum principle and the standard H_∞ problem*, Math. Control Systems Signals, 3 (1989), pp. 301–324.

[16] ———, *The standard H_∞ problem and the maximum principle: the general linear case,* Tech. Report 192, University of Texas, Dallas, TX, May 1989; in Proc. 28th IEEE Conference on Decision and Control, December 1989, pp. 403–406.

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

[18] G. ZAMES, *Feedback and optimal sensitivity : Model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, 26 (1981), pp. 301–320.

[19] K. ZHOU AND P. P. KHARGONEKAR, *An algebraic Riccati equation approach to H_∞ optimization*, Systems Control Lett., 11 (1988), pp. 85-92.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 6, pp. 1394–1413, November 1991

007

H^∞ CONTROL OF LINEAR TIME-VARYING SYSTEMS: A STATE-SPACE APPROACH*

R. RAVI[†], K. M. NAGPAL[‡], AND P. P. KHARGONEKAR[§]

Abstract. In this paper the standard problem of H^∞ control theory for finite-dimensional linear time-varying continuous-time plants is considered. The problem is: given a real number $\gamma > 0$, find (if one exists) an internally stabilizing controller such that the closed-loop operator norm is less than γ. Under rather weak assumptions on the plant model, it is shown that a solution to this problem exists if and only if a pair of matrix Riccati differential equations admits positive semidefinite stabilizing solutions. State-space formulae for one solution to the problem are also given.

Key words. H_∞ control theory, linear time-varying systems, Riccati differential equations, optimal control

AMS(MOS) subject classifications. 93B50, 93C35, 93C05, 49A40

1. Introduction.

After the introduction of the H^∞ control problem by Zames [24], initial developments in H^∞ control theory were based on frequency domain and operator theoretic methods. The book by Francis [9] contains an excellent account of the results obtained using this approach. While most of the results were for linear time-invariant systems, several results on H^∞ or uniformly optimal control of linear time-varying discrete-time systems were obtained using these methods. See, for example, [6], [7], [8], [10], and the references cited there. Recently, some new and very interesting results on the H^∞ control of slowly varying systems using operator theoretic methods have also been obtained [23].

A major new development in H^∞ control theory during the last three years has been the introduction of state-space methods. (State-space representations were used even earlier in computing solutions to H_∞ control problems. By *state-space methods*, we are referring to the systematic use of state-space ideas, e.g., state-feedback, state-estimation, separation principle, etc., for deriving and computing solutions.) This state-space approach has proved to be quite successful in providing simple and intuitive solutions to the H^∞ control problem. The interested reader is referred to [5], [13] and the references contained there for these recent developments.

The state-space approach is particularly natural and appealing for the problem of H^∞ control of linear time-varying systems. Tadmor [21], [20] was the first to apply the state-space approach for linear time-varying systems. Recently, a simple game theoretic solution to the H^∞ control problem for the finite horizon case was given in [15]. In [21], Tadmor gave a solution to the H^∞ control problem for the finite horizon case, while [20] contained a solution for the infinite horizon case. It should be noted that the infinite horizon case is significantly more difficult since stability becomes an

* Received by the editors April 16, 1990; accepted for publication (in revised form) September 10, 1990. This work was supported in part by National Science Foundation grant ECS-9096109, Air Force Office of Scientific Research contract AFOSR-90-0053, and Army Research Office grant DAAL03-90-G-0008. An abridged version of this paper has appeared in the Proceedings of the 1990 IEEE Conference on Decision and Control.

[†] Department of Electrical Engineering, University of Minnesota, Minneapolis, Minnesota, 55455. Present address, Control Systems Laboratory, GE-CRD, Schenectady, New York. This author was also supported by a University of Minnesota Graduate Dissertation Fellowship.

[‡] Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan, 48109. Present address, Coordinated Science Laboratory, University of Illinois, Urbana, Illinois.

[§] Department of Electrical Engineering and Computer Science, The University of Michigan, Ann Arbor, Michigan 48109.

important issue, whereas in the finite horizon case this is completely absent. Tadmor derived his results under the hypothesis that the input matrix for the exogenous signal and the output matrix for the controlled variables are nonsingular. The results in [15], for the finite horizon case, are derived under much less restrictive assumptions.

In the present paper, we give a solution to the infinite horizon case building on the previous work of [5], [15], [21], and [20]. We obtain necessary and sufficient conditions for the existence of solutions under the assumption that the plant is stabilizable from the exogenous inputs and detectable from the controlled outputs. These assumptions are significantly weaker than those in [20] — thus the main contribution of the paper is the generalization of existing results in [20] by relaxing assumptions to stabilizability and detectability. While it is trivial to conjecture our main results in view of the results in [5], the proofs of these results are not so easy to generalize. Indeed, even in the context of the linear quadratic Gaussian (LQG) problem, such generalizations involve significant technical difficulties. See (for the discrete-time case) [2] and [3] in this connection. For the H_∞ control problem, these generalizations are nontrivial as well. Indeed, our proof techniques are quite different, and in our opinion simpler, than those employed in [20]. Thus it is hoped that, in addition to generalizing existing results, these techniques will provide an alternative approach to proving some of the existing results even in the linear time-invariant case.

This paper is organized as follows. In §2, we set up the notation and the problem formulation. Here we present a few preliminary results and some of the machinery needed for proving the main result. Section 3 contains the statement and a proof of the main result.

2. Notation and preliminary results. Let $\mathcal{R}_+$ (respectively, $\mathcal{R}_-$) be the set of nonnegative (respectively, nonpositive) real numbers, and let $\mathcal{R} = \mathcal{R}_+ \cup \mathcal{R}_-$. Let $\mathcal{L}_2^m(\mathcal{R}_+)$ be the space of square integrable functions on $\mathcal{R}_+$ with values in $\mathcal{R}^m$. The extended space is defined as $\mathcal{L}_{2,e}^m(\mathcal{R}_+) := \{f : f$ is a measurable function on $\mathcal{R}_+$, $P_t f \in \mathcal{L}_2^m(\mathcal{R}_+)$, for all $t \in \mathcal{R}_+\}$ where $(P_t f)(\tau) = f(\tau)$ if $\tau \le t$, and 0 if $\tau > t$. In the sequel, the notation for these spaces will be abbreviated to $\mathcal{L}_2$ and $\mathcal{L}_{2,e}$, respectively, when the dimensions are unambiguous.

Consider the linear system

$$\Sigma_G := \left\{ \begin{array}{rcl} \dot{x}(t) & = & A(t)x(t) + B(t)u(t), \quad x(t_0) = x_0, \\ y(t) & = & C(t)x(t) + D(t)u(t). \end{array} \right. \tag{1}$$

Here $x(t)$, $u(t)$, and $y(t)$, when evaluated at time t, belong to $\mathcal{R}^n$, $\mathcal{R}^m$, and $\mathcal{R}^p$, respectively, and the matrices are all bounded functions of t. With $x(0) = 0$ it can be shown that Σ_G generates a causal, linear operator G mapping $\mathcal{L}_{2,e}^m(\mathcal{R}_+)$ to $\mathcal{L}_{2,e}^p(\mathcal{R}_+)$ given by

$$y(t) = \int_0^t C(t)\Phi_G(t,\tau)B(\tau)u(\tau)d\tau + D(t)u(t), \tag{2}$$

where $\Phi_G(t,\tau)$ is the state transition matrix of the homogeneous part of (1). In terms of Σ_G we have the following definitions.

DEFINITION. The system Σ_G is said to be *exponentially* (*or internally*) *stable* if there exist $c_1, c_2 > 0$ subject to $||\Phi_G(t,\tau)|| \le c_1 e^{-c_2(t-\tau)}$, for all $t \ge \tau$.

As this definition of stability deals with the homogeneous part of Σ_G alone, we will use the phrase, "A is stable," to mean that Σ_G is exponentially stable. By the same reasoning, we will also use the notation Σ_A interchangeably with Σ_G when referring to the state transition matrix.

If Σ_G is exponentially stable, then the input-output operator $G : \mathcal{L}_2^m(\mathcal{R}_+) \to \mathcal{L}_2^p(\mathcal{R}_+)$ is a bounded linear operator and its induced norm is defined as $\|G\| := \sup_{\|u\|\neq 0}(\|y\|/\|u\|)$. Note that an alternate notation for the input-output operator mapping u to y is T_{yu}.

DEFINITION. The system Σ_G is said to be *stabilizable* (respectively, *detectable*) if there exists a bounded function $K(t)$ (respectively, $L(t)$) such that the system $\dot{x}(t) = (A - BK)(t)x(t)$ (respectively, $\dot{x}(t) = (A - LC)(t)x(t)$) is exponentially stable. We also use the notation (A, B) stabilizable (respectively, (A, C) detectable) to denote this.

It is a fact that if Σ_G is stabilizable and detectable then it is exponentially stable if and only if G is a bounded operator (a proof for the discrete-time case is in [1]; the continuous-time case is also easily shown).

We briefly review the concept of a dual, which was first introduced by Kalman (see, e.g., [12]) as a means of relating regulation and estimation. We begin by considering the system Σ_G defined only over a finite time interval $[0, T]$. Let $t^* := -t$ and define $A(t^*) := A(t)$, $B(t^*) := B(t)$, $C(t^*) := C(t)$, and $D(t^*) := D(t)$. Then

$$(3) \qquad \Sigma_{G^*} := \begin{cases} \dot{x}^*(t^*) &= A'(t^*)x^*(t^*) + C'(t^*)u^*(t^*), \qquad x^*(-T) = x_0, \\ y^*(t^*) &= B'(t^*)x^*(t^*) + D'(t^*)u^*(t^*) \end{cases}$$

is called the *dual* of Σ_G. It can be shown that if $x^*(-T) = 0$ then Σ_{G^*} generates a causal, linear, and bounded system G^* mapping $\mathcal{L}_2^p([-T, 0])$ to $\mathcal{L}_2^m([-T, 0])$. Moreover, G^* can be identified with the *adjoint* of G. We are mainly interested in the case when $T = \infty$, and here we must make the additional assumption that the original system Σ_G is exponentially stable. Let $u^* \in \mathcal{L}_2^p(\mathcal{R}_+)$ and define

$$(4) \qquad x^*(t) := \int_t^\infty \Phi_G'(\tau, t)C'(\tau)u^*(\tau)d\tau.$$

By differentiating this we find that $x^*(t)$ satisfies

$$(5) \qquad -\dot{x}^*(t) = A'(t)x^*(t) + C'(t)u^*(t).$$

Because Σ_G is exponentially stable it can be shown, by applying Schwartz's inequality to (4), that $\lim_{t\to\infty} x^*(t) = 0$. If we define the output $y^*(t)$ as

$$(6) \qquad y^*(t) = B'(t)x^*(t) + D'(t)u^*(t),$$

then it is clear that $< u^*, y > = < y^*, u >$, for all $u \in \mathcal{L}_2^m(\mathcal{R}_+)$ and for all $u^* \in \mathcal{L}_2^p(\mathcal{R}_+)$, and $y = Gu$. It follows that (5) and (6) together define the adjoint of G. If in (5) and (6) we further make the change of variable $t^* = -t$, we obtain Σ_{G^*} exactly as in (3) but with the boundary condition replaced by $\lim_{t^*\to-\infty} x^*(t^*) = 0$. Note that Σ_{G^*} is defined only on $\mathcal{R}_-$ and is stable in the sense that there exist $c_1, c_2 > 0$ subject to $\|\Phi_{G^*}(t^*, \tau^*)\| \leq c_1 e^{-c_2(t^* - \tau^*)}$, and all $\tau^* \leq t^*$, for all $t^* \in \mathcal{R}_-$.

We are now ready to formulate the control problem considered in this paper. Consider the control system shown in Fig. 1. Let the finite-dimensional linear time-varying (FDLTV) plant G be described by

$$(7) \qquad \Sigma_G := \begin{cases} \dot{x}(t) &= A(t)x(t) + B_1(t)w(t) + B_2(t)u(t), \\ z(t) &= C_1(t)x(t) + D_{12}(t)u(t), \\ y(t) &= C_2(t)x(t) + D_{21}(t)w(t), \end{cases}$$

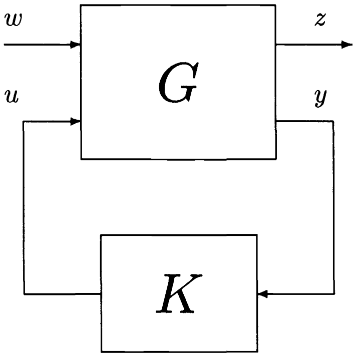


FIG. 1. *Standard feedback configuration.*

or in more compact packed matrix notation

$$\Sigma_G := \left[\begin{array}{c|cc} A & B_1 & B_2 \\ \hline C_1 & 0 & D_{12} \\ C_2 & D_{21} & 0 \end{array}\right].$$

Note that we have assumed that the direct feedthrough terms D_{11} and D_{22} are zero. This assumption, especially the first one, greatly reduces the length and complexity of the formulae to be derived in §3. The general forms can be obtained by constructions similar to those in [15], [17], and [25].

Let T_{zw} denote the closed-loop operator mapping w to z. Then the standard problem of H^∞ control theory for linear time-varying systems can be stated as follows. Given the FDLTV system G with a realization Σ_G, and a real number $\gamma > 0$, find (when one exists) an FDLTV controller K that exponentially (internally) stabilizes the closed-loop system and makes $||T_{zw}|| < \gamma$. Any controller that satisfies the above condition will be called an *admissible* controller.

Note that we are only seeking a controller that is *suboptimal* relative to the number γ. For most cases this is sufficient because we can come as close to the optimal value, $\inf_K ||T_{zw}||$, as we want by iterating on γ.

We make the following simplifying assumptions on Σ_G:

(A1) $D_{12}'(t)(C_1(t)\ \ D_{12}(t)) = (0\ \ I)$.

(A2) $\begin{pmatrix} B_1(t) \\ D_{21}(t) \end{pmatrix} D_{21}'(t) = \begin{pmatrix} 0 \\ I \end{pmatrix}$.

(A3) (A, B_1) is stabilizable.

(A4) (A, C_1) is detectable.

(A5) (A, B_2) is stabilizable.

(A6) (A, C_2) is detectable.

The full rank conditions on D_{12} and D_{21} are restrictive but essential for the theory to work. To remove this assumption we can use a reasoning similar to that in [14], where it is shown that if these conditions are not satisfied, we can perturb the matrices slightly so that they are satisfied, ensuring at the same time that the solution to the new problem is also a solution to the original one. The crucial fact that allows us to use this argument is that we are only dealing with a suboptimal problem. Dealing with these assumptions in a more direct manner is more difficult and has been done in the *time-invariant* case via the use of methods from singular control theory and almost invariant subspaces [19], [18]. It is unclear at this time if these methods can be

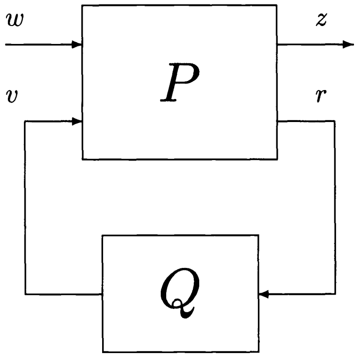


FIG. 2. *Feedback configuration for Lemma* 2.2.

extended to linear time-varying systems since very little is known about the singular cases even for the LQG problem for linear time-varying systems.

Assumptions (A5), (A6) are necessary for the existence of an exponentially stabilizing controller. For LTI systems, assumptions (A3), (A4) can be relaxed further to requiring that certain rank conditions be satisfied on the imaginary axis [11]. In the linear time-varying case, however, it is unclear what the corresponding assumptions should be since the role of the imaginary axis is very strongly related to time-invariance. This is also related to whether we consider linear time-varying systems on the half line or the entire real line for the time-axis. Techniques from operator theory suggest that this is related to the existence of certain inner-outer factorizations. We will restrict ourselves to assumptions (A3), (A4) in this paper, leaving further extensions for future work.

Before going on to the main result, we state three preliminary lemmas that we will need in the proof of Theorem 3.1. The first is a version of the Lyapunov stability theorem proved for discrete time systems in [2]. The following continuous-time version is proved in [16].

LEMMA 2.1. *Let Σ be defined as $\dot{x}(t) = A(t)x(t)$ and let C and B be such that (A, C) and (A, B) are detectable and stabilizable, respectively. Then either of the following statements is a sufficient condition for the system Σ to be exponentially stable.*

(1) *There exists a bounded nonnegative definite function $X(t)$, $t \in [0, \infty)$ such that*

$$\dot{X}(t) + A'(t)X(t) + X(t)A(t) = -C'(t)C(t).$$

(2) *There exists a bounded nonnegative definite function $Y(t)$, $t \in [0, \infty)$ such that*

$$\dot{Y}(t) - A(t)Y(t) - Y(t)A'(t) = B(t)B'(t).$$

The next result is the time-varying version of Lemma 15 in [5].

LEMMA 2.2. *Let P be a system partitioned as follows:*

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix}$$

and let Q be another system connected to P as in Fig. 2. Suppose that P has an exponentially stable realization and that Q is given by a stabilizable and detectable *state-space realization. In addition, suppose that the resulting realization for the closed loop system is also stabilizable and detectable (from w and z, respectively). Let P be isometric, (i.e., $\|w\|^2+\|v\|^2 = \|z\|^2+\|r\|^2$, for all $v, w \in \mathcal{L}_2$), and let P_{21}^{-1} exist and be stable. Then the closed-loop system is exponentially stable and $||T_{zw}|| < 1$ if and only if Q is exponentially stable and $||Q|| < 1$, where T_{zw} is the closed-loop input-output operator mapping w to z.*

Proof. Sufficiency. As $||Q|| < 1$ and $||P_{22}|| \leq 1$ it follows that $(I-P_{22}Q)^{-1}$ exists as a stable operator. This, along with the fact that both P and Q are stable, implies closed-loop stability of the system in Fig. 2. Internal exponential stability then follows from the hypothesis that the closed-loop system is stabilizable and detectable. The norm bound is also easy to establish. From the isometric property of P it follows that

$$||z||^2 + ||r||^2 = ||w||^2 + ||v||^2. \tag{8}$$

As $||Q|| < 1$ we have

$$\begin{aligned} ||v||^2 - ||r||^2 &\leq -\varepsilon ||r||^2 \\ &\leq -\frac{\varepsilon}{||T_{wr}||^2} ||w||^2 \end{aligned} \tag{9}$$

for some $\varepsilon > 0$ and where $T_{wr} := P_{21}^{-1} - P_{21}^{-1}P_{22}Q$ is a bounded operator. From (8) and (9) it clearly follows that $||T_{zw}|| < 1$.

Necessity. From Fig. 2 we have

$$\begin{aligned} r &= P_{21}w + P_{22}v \\ &= P_{21}w + P_{22}Qr \\ \Rightarrow w &= P_{21}^{-1}r - P_{21}^{-1}P_{22}Qr. \end{aligned} \tag{10}$$

Let $\mathcal{M} := \{r \in \mathcal{L}_2 \text{ subject to } Qr \in \mathcal{L}_2\}$. From (10) it is easy to see that whenever $r \in \mathcal{M}$ we have $w \in \mathcal{L}_2$. This means that given any signal $r \in \mathcal{M}$ there exists a $w \in \mathcal{L}_2$ that generates it as an input to Q. From (8) and the fact that $||T_{zw}|| < 1$, it follows that whenever $r \in \mathcal{M}$,

$$||v||^2 - ||r||^2 = ||z||^2 - ||w||^2 \leq -\epsilon ||w||^2 \tag{11}$$

for some $\epsilon > 0$. Furthermore, as the closed loop is stable we must have that T_{rw}, the closed-loop operator mapping w to r, is bounded. Along with (11) this implies that

$$||v||^2 - ||r||^2 \leq -\frac{\epsilon}{||T_{rw}||} ||r||^2,$$

from which it follows that

$$\sup_{r \in \mathcal{M}, r \neq 0} \frac{||Qr||}{||r||} < 1. \tag{12}$$

To complete the proof we need to show that Q is stable. If we show this then (12), together with the definition of $\mathcal{M}$, implies that $||Q|| < 1$.

To establish the stability of Q we use a time-varying version of the proof of Theorem 1 in [22]. Assume Q is unstable. Because Q has finite gain on the set $\mathcal{M}$ (see (12)), it follows from Lemma 1 in [22] that $\mathcal{M}$ is a closed subspace of $\mathcal{L}_2$. If $M \neq \mathcal{L}_2$ then there is at least one nonzero d in $\mathcal{M}^\perp$, the orthogonal complement of

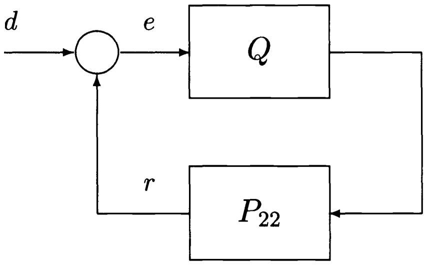


FIG. 3.

$\mathcal{M}$ in $\mathcal{L}_2$. As the closed loop is stable, it follows that the feedback system in Fig. 3 is also stable, and so for any $d \in \mathcal{L}_2$ both e and r are in $\mathcal{L}_2$, and, furthermore, we have $e \in \mathcal{M}$. Choose $d \in \mathcal{M}^\perp$. Then

$$\begin{aligned} & r = e - d \\ & \Rightarrow ||r||^2 = ||e||^2 + ||d||^2 \quad (\text{since } e \in \mathcal{M} \text{ and } d \in \mathcal{M}^\perp) \\ (13) \qquad & \Rightarrow ||e|| < ||r||. \end{aligned}$$

On the other hand, because $||P_{22}|| \leq 1$ and $||Q|| < 1$, it follows that $||e|| > ||r||$, which is a contradiction of (13). This proves our claim. □

Finally, we state a consequence of the definition of detectability that we will need in the following. Consider the system

$$(14) \qquad \Sigma := \begin{cases} \dot{x}(t) = A(t)x(t) + B(t)u(t), & x(0) = 0, \\ z(t) = C(t)x(t), \end{cases}$$

and assume that Σ is detectable.

LEMMA 2.3. *Let the system Σ be as in* (14). *If Σ is detectable then there exists a constant $\zeta > 0$ such that for any $T > 0$ and a reachable state x_T at time $t = T$,*

$$(15) \qquad \inf_{u \in \mathcal{L}_2([0,T])} (||P_T z||^2 + ||P_T u||^2 : \; x(0) = 0 \text{ and } x(T) = x_T) \geq \zeta ||x_T||^2.$$

Proof. As the system is detectable, there is a bounded function $L(t)$ subject to $(A - LC)(t)$ that is exponentially stable. System (14) can be equivalently written as

$$\begin{aligned} \dot{x}(t) &= (A - LC)(t)x(t) + B(t)u(t) + L(t)z(t), \qquad x(0) = 0, \\ z(t) &= C(t)x(t). \end{aligned}$$

As $(A - LC)(t)$ is exponentially stable, clearly there exists a number $\rho < \infty$ (independent of T) subject to

$$||x(T)||^2 < \rho(||P_T z||^2 + ||P_T u||^2).$$

The conclusion follows immediately by taking $\zeta = 1/\rho$. □

3. Main result.

THEOREM 3.1. *Let Σ_G satisfy assumptions* (A1)–(A6). *Then there exists a controller K that exponentially stabilizes the system Σ_G and makes $||T_{zw}|| < \gamma$ if and only if*

(1) *There is a bounded nonnegative definite solution to the Riccati equation*

$$\begin{aligned} -\dot{X}_\infty(t) &= A'(t)X_\infty(t) + X_\infty(t)A(t) \\ &\quad -X_\infty(t)(B_2(t)B_2'(t) - \tfrac{1}{\gamma^2}B_1(t)B_1'(t))X_\infty(t) + C_1'(t)C_1(t) \end{aligned} \tag{16}$$

subject to the system $\dot{x}(t) = (A-(B_2B_2'-\gamma^{-2}B_1B_1')X_\infty)(t)x(t)$ *is exponentially stable, and*

(2) *There is a bounded nonnegative definite solution to the Riccati equation*

$$\begin{aligned} \dot{Y}_{tmp}(t) &= A_{tmp}(t)Y_{tmp}(t) + Y_{tmp}(t)A_{tmp}'(t) \\ &\quad -Y_{tmp}(t)(C_2'(t)C_2(t) - \tfrac{1}{\gamma^2}X_\infty(t)B_2(t)B_2'(t)X_\infty(t))Y_{tmp}(t) \\ &\quad +B_1(t)B_1'(t), \quad Y_{tmp}(0) = 0 \end{aligned} \tag{17}$$

subject to the system $\dot{x}(t) = (A_{tmp} - Y_{tmp}(C_2'C_2 - \gamma^{-2}X_\infty B_2B_2'X_\infty))(t)x(t)$ *is exponentially stable, where*

$$A_{tmp} := \left(A + \tfrac{1}{\gamma^2}B_1B_1'X_\infty\right). \tag{18}$$

If these two conditions are met, then an admissible controller is given by

$$\Sigma_K := \begin{cases} \dot{q}(t) = (A - (B_2B_2' - \gamma^{-2}B_1B_1')X_\infty - Y_{tmp}C_2'C_2)(t)q(t) \\ \qquad\quad +Y_{tmp}(t)C_2'(t)y(t), \\ u(t) = -B_2'(t)X_\infty(t)q(t). \end{cases} \tag{19}$$

Note that in (16), as opposed to (17), no boundary condition is specified. This is similar to the corresponding results for the classical LQG problem in the infinite horizon case where the control Riccati equation does not have a boundary condition. It will be seen in the proof that X_∞ is obtained as the limit of the solutions of finite horizon Riccati differential equations as the terminal time approaches infinity.

We have given the necessary and sufficient conditions here in terms of two coupled Riccati differential equations. Since the coupling is only one way, i.e., the solution to the first enters the second but not vice versa, this presents no problems. It is possible to give an alternative set of necessary and sufficient conditions involving two uncoupled Riccati differential equations and a spectral radius condition. This is quite standard; see, for example, [15] and [20].

Proof. Note that there is no loss of generality in replacing γ by 1—hence in the proof we will work with this simplifying normalization.

Sufficiency. We will show that K, defined in (19), both stabilizes the system and makes $||T_{zw}|| < 1$. We start by making the change of variables $v := u + B_2'X_\infty x$ and $r := w - B_1'X_\infty x$. In terms of these variables we have a new system P defined by

$$\Sigma_P := \begin{cases} \dot{x} = (A - B_2B_2'X_\infty)x + B_1w + B_2v, \\ z = (C_1 - D_{12}B_2'X_\infty)x + D_{12}v, \\ r = -B_1'X_\infty x + w \end{cases} \tag{20}$$

and a corresponding controller C defined by

$$\Sigma_C := \begin{cases} \dot{e} = (A_{tmp} - Y_{tmp}C_2'C_2)e + (B_1 - Y_{tmp}C_2'D_{21})r, \\ v = B_2'X_\infty e, \end{cases} \tag{21}$$

where $e := x - q$.

We will now show that P is stable and norm preserving, i.e., $||(z' \ r')'|| = ||(w' \ v')'||$, for all $(w, v) \in \mathcal{L}_2 \times \mathcal{L}_2$. To establish this, let us rewrite (16) as

$$\begin{aligned}&\dot{X}_\infty(t) + (A - B_2B_2'X_\infty)'(t)X_\infty(t) + X_\infty(t)(A - B_2B_2'X_\infty)(t)\\&= -(X_\infty B_2B_2'X_\infty)(t) - (X_\infty B_1B_1'X_\infty)(t) - (C_1'C_1)(t).\end{aligned}$$

As (A, C_1) is detectable it follows that $((A - B_2B_2'X_\infty), (C_1' \mid X_\infty B_2 \mid X_\infty B_1)')$ is detectable; therefore, by Lemma 2.1, we have that $\dot{x}(t) = (A - B_2B_2'X_\infty)(t)x(t)$ is exponentially stable. To establish the norm condition, we differentiate $x'(t)X_\infty(t)x(t)$ along the trajectory of Σ_P to get

$$\begin{aligned}\frac{d(x'X_\infty x)}{dt} &= \dot{x}'(t)X_\infty(t)x(t) + x'(t)\dot{X}_\infty(t)x(t) + x'(t)X_\infty(t)\dot{x}(t)\\&= ((A - B_2B_2'X_\infty)(t)x(t) + B_1(t)w(t) + B_2(t)v(t))'X_\infty(t)x(t)\\&\quad -x'(t)(A'X_\infty + X_\infty A - X_\infty(B_2B_2' - B_1B_1')X_\infty + C_1'C_1)(t)x(t)\\&\quad +x'(t)X_\infty(t)((A - B_2B_2'X_\infty)(t)x(t) + B_1(t)w(t) + B_2(t)v(t)).\end{aligned}$$

Expanding this, substituting from (20) and (21), and cancelling a number of terms, gives

$$\frac{d(x'X_\infty x)}{dt} = w'w(t) - r'r(t) + v'v(t) - z'z(t). \tag{22}$$

By assumption, $x(0) = 0$. Furthermore, we have established that the system Σ_P is exponentially stable, which means that for every $(w, v) \in \mathcal{L}_2 \times \mathcal{L}_2$ we have that $\lim_{t\to\infty} x(t) = 0$. Consequently, we can integrate (22) between 0 and ∞ to get

$$0 = ||w||^2 - ||r||^2 + ||v||^2 - ||z||^2$$

or

$$||w||^2 + ||v||^2 = ||r||^2 + ||z||^2.$$

Let us now partition P as in Lemma 2.2. As $\dot{x}(t) = (A-(B_2B_2'-B_1B_1')X_\infty)(t)x(t)$ is exponentially stable, it follows that P_{21}^{-1} exists and is stable. Next we establish that Σ_C is stable and that $||C|| < 1$. We rewrite (17) as

$$\begin{aligned}&\dot{Y}_{tmp}(t) - (A_{tmp} - Y_{tmp}C_2'C_2)(t)Y_{tmp}(t) - Y_{tmp}(t)(A_{tmp} - Y_{tmp}C_2'C_2)'(t)\\&= (Y_{tmp}C_2'C_2Y_{tmp})(t) + (Y_{tmp}X_\infty B_2B_2'X_\infty Y_{tmp})(t) + (B_1B_1')(t).\end{aligned} \tag{23}$$

By assumption we have that $\dot{x}(t) = (A_{tmp} - Y_{tmp}C_2'C_2 + Y_{tmp}X_\infty B_2B_2'X_\infty)(t)x(t)$ is exponentially stable, which means $((A_{tmp} - Y_{tmp}C_2'C_2), (Y_{tmp}X_\infty B_2))$ and, consequently, $((A_{tmp} - Y_{tmp}C_2'C_2), (Y_{tmp}C_2' \mid Y_{tmp}X_\infty B_2 \mid B_1))$ is stabilizable. It follows from Lemma 2.1 that Σ_C is exponentially stable. Next we show that $||C|| < 1$. Let us define the dual Σ_{C^*} as

$$\begin{aligned}\dot{e}^*(t^*) &= (A_{tmp}' - C_2'C_2Y_{tmp})(t^*)e^*(t^*) + (X_\infty B_2)(t^*)r^*(t^*),\\ v^*(t^*) &= (B_1' - D_{21}'C_2Y_{tmp})(t^*)e^*(t^*),\end{aligned} \tag{24}$$

where $Y_{tmp}(t^*)$ is the solution to the time-reversed Riccati equation obtained from (17) as

$$\begin{aligned}-\dot{Y}_{tmp}(t^*) &= A_{tmp}(t^*)Y_{tmp}(t^*) + Y_{tmp}(t^*)A_{tmp}'(t^*) + B_1B_1'(t^*)\\&\quad -Y_{tmp}(t^*)(C_2'C_2 - X_\infty B_2B_2'X_\infty)(t^*)Y_{tmp}(t^*), \qquad Y_{tmp}(0) = 0.\end{aligned} \tag{25}$$

We consider the function $e^{*\prime} Y_{tmp} e^*(t^*)$ and calculate its derivative along the trajectory of (24). Substituting for $Y_{tmp}(t^*)$ from (25) gives us

(26)
$$\begin{aligned}\frac{d(e^{*\prime} Y_{tmp} e^*)}{dt^*} &= -v^{*\prime} v^*(t^*) + r^{*\prime} r^*(t^*) \\ &\quad -(r^* - B_2' X_\infty Y_{tmp} e^*)'(r^* - B_2' X_\infty Y_{tmp} e^*)(t^*).\end{aligned}$$

As we have established that Σ_{C^*} is exponentially stable, we have $\lim_{t^* \to -\infty} e^*(t^*) = 0$. This coupled with the fact that $Y_{tmp}(0) = 0$ allows us to integrate (26) from $-\infty$ to 0 to get

$$0 = ||r^*||^2 - ||v^*||^2 - ||r^* - B_2' X_\infty Y_{tmp} e^*||^2$$

or

$$||v^*||^2 - ||r^*||^2 = -||r^* - B_2' X_\infty Y_{tmp} e^*||^2. \qquad (27)$$

Clearly, $r^* - B_2' X_\infty Y_{tmp} e^*$ is the output of the following system with input r^*:

$$\left[\begin{array}{c|c} A_{tmp}' - C_2' C_2 Y_{tmp} & X_\infty B_2 \\ \hline -B_2' X_\infty Y_{tmp} & I \end{array}\right].$$

Note that the inverse of this system is given by

$$\left[\begin{array}{c|c} A_{tmp}' - C_2' C_2 Y_{tmp} + X_\infty B_2 B_2' X_\infty Y_{tmp} & X_\infty B_2 \\ \hline B_2' X_\infty Y_{tmp} & I \end{array}\right],$$

which maps $r^* - B_2' X_\infty Y_{tmp} e^*$ to r^*. As this is a stable system there exist δ with $0 < \delta < \infty$ such that $||r^*||^2 \leq \delta ||r^* - B_2' X_\infty Y_{tmp} e^*||^2$ or $-||r^* - B_2' X_\infty Y_{tmp} e^*||^2 \leq -1/\delta ||r^*||^2$. Coupling this with (27) gives

$$||v^*||^2 - ||r^*||^2 \leq -\frac{1}{\delta} ||r^*||^2, \qquad (28)$$

which means that $||C^*|| < 1$ or equivalently that $||C|| < 1$. A small amount of algebra shows that P and C satisfy all the assumptions of Lemma 2.2 (when C is substituted for Q). Therefore, as Σ_C is exponentially stable and $||C|| < 1$ it follows that the closed-loop system is stable and $||T_{zw}|| < 1$.

Necessity. In this part we assume that there exists an admissible controller; i.e., we have a controller K that stabilizes the closed loop and, in addition, makes $||T_{zw}|| < 1$, and show that the two Riccati differential equations have stabilizing solutions. For each case we will show that a nonnegative solution exists, that it is bounded, and, finally, that it is stabilizing.

Existence and boundedness of X_∞. Let us begin with a finite time horizon solution to the Riccati equation, which we will then extend to the infinite horizon. Let $T < \infty$ be the terminal time. As K is a *causal* admissible controller over $\mathcal{R}_+$ it is also admissible over $[0, T]$ for any T. In other words,

$$\begin{aligned} & ||T_{zw}|| < 1 \\ \Rightarrow & \sup_{w \neq 0} \frac{||z||}{||w||} < 1 \\ \Rightarrow & \sup_{P_T w \neq 0} \frac{||P_T z||}{||P_T w||} < 1 \quad \forall T \in \mathcal{R}_+. \end{aligned} \qquad (29)$$

From existing results on the finite horizon H_∞ control of linear time-varying systems in [21], [15], we conclude that this condition is sufficient for the existence of a bounded nonnegative definite solution to the finite horizon Riccati equation,

$$-\dot{X}_T(t) = A'(t)X_T(t) + X_T(t)A(t) - X_T(t)(B_2B_2' - B_1B_1')(t)X_T(t) + C_1'(t)C_1(t), \tag{30}$$
$$X_T(T) = 0.$$

The existence of a solution to (30) implies certain properties for the following differential game. Consider the system

$$\begin{array}{rcll} \dot{x}(t) & = & A(t)x(t) + B_1(t)w(t) + B_2(t)u(t), & x(0) = x_0, \\ z(t) & = & C_1(t)x(t) + D_{12}(t)u(t). & \end{array} \tag{31}$$

Set T to be the terminal time and define the cost function over $[0, T]$ as $J_T(w, u) := ||P_Tz||^2 - ||P_Tw||^2$. The two opposing players are the control input u and the exogenous signal w and we assume that these take values in $\mathcal{L}_2[0, T]$. The objective of the control input u is to minimize the cost while that of the exogenous signal w is to maximize it. Define $u_T^\star(x)(t) := -B_2'(t)X_T(t)x(t)$ and $w_T^\star(x)(t) := B_1'(t)X_T(t)x(t)$. Then a standard calculation involving completion of squares shows that for any u and w in $\mathcal{L}_2[0, T]$

$$J_T(w, u_T^\star(x)) \leq J_T(w_T^\star(x), u_T^\star(x)) = x_0'X_T(0)x_0 \leq J_T(w_T^\star(x), u). \tag{32}$$

It is important to note that the strategies $u_T^\star$ and $w_T^\star$ depend on x and are *not* open-loop saddle point strategies.

Now consider two different terminal times T_1 and T_2 with $T_1 \leq T_2$. We will next show that $X_{T_1}(t) \leq X_{T_2}(t)$, for all $t \in [0, T_1]$. Consider, first, the game over $[0, T_1]$. Suppose $w = w_{T_1}^\star(x)$. Then from (32) we have

$$x_0'X_{T_1}(0)x_0 \leq J_{T_1}(w_{T_1}^\star(x), u) \tag{33}$$

for any $u \in \mathcal{L}_2[0, T_1]$. Now let the terminal time be T_2 and let w_{T_2} be defined on $[0, T_2]$ as

$$w_{T_2}(t) := \begin{cases} w_{T_1}^\star(x)(t), & t \leq T_1, \\ 0, & T_1 < t \leq T_2. \end{cases}$$

Clearly,

$$\begin{aligned} J_{T_2}(w_{T_2}, u) &= \int_0^{T_1} (||z(t)||^2 - ||w_{T_2}(t)||^2)dt + \int_{T_1}^{T_2} (||z(t)||^2 - ||w_{T_2}(t)||^2)dt \\ &= J_{T_1}(w_{T_1}^\star(x), u) + \int_{T_1}^{T_2} ||z(t)||^2 dt \\ &\geq J_{T_1}(w_{T_1}^\star(x), u) \\ &\geq J_{T_1}(w_{T_1}^\star(x), u_{T_1}^\star(x)) = x_0'X_{T_1}(0)x_0. \end{aligned} \tag{34}$$

The above chain of inequalities holds for any $u \in \mathcal{L}_2[0, T_2]$, and, in particular, it holds for $u = u_{T_2}^\star(x)$. On the other hand, from (32) it follows that

$$x_0'X_{T_2}(0)x_0 = J_{T_2}(w_{T_2}^\star(x), u_{T_2}^\star(x)) \geq J_{T_2}(w_{T_2}, u_{T_2}^\star(x)). \tag{35}$$

Combining (34) and (35) we get that $x_0'X_{T_2}(0)x_0 \geq x_0'X_{T_1}(0)x_0$; and as this is true for any $x_0 \in \mathcal{R}^n$ it follows that $X_{T_2}(0) \geq X_{T_1}(0)$. But there is nothing sacred about choosing the starting time as $t = 0$. We could have played this game starting at any time $t \in [0, T_1]$ and hence it follows that $X_{T_2}(t) \geq X_{T_1}(t)$ for all $t \in [0, T_1]$.

Next we show that $X_T(t)$ remains a bounded function of time for every T. As K is admissible it follows, from (29), that (with $x_0 = 0$) there exist $\epsilon > 0$ subject to $||P_Tz||^2 - ||P_Tw||^2 \leq -\epsilon||P_Tw||^2$, for all $w \in \mathcal{L}_2([0,T])$ and all $T \in R_+$. Next, let $x_0 \neq 0$, let $u = Ky$, and set the initial states of the controller to be zero. As K is linear the output z of (31) to any input w can be written as

$$z(t) = z_1(t) + z_2(t), \tag{36}$$

where $z_1(t)$ is the homogeneous part of the solution (depending only on x_0) and $z_2(t)$ is the forced part (depending only on w). Clearly, we have, for any finite T,

$$||P_Tz||^2 \leq ||P_Tz_1||^2 + ||P_Tz_2||^2 + 2||P_Tz_1||\,||P_Tz_2||. \tag{37}$$

However, because K is an admissible controller the resulting closed-loop system is exponentially stable, which further implies that both z_1 and z_2 are in $\mathcal{L}_2(\mathcal{R}_+)$ for any $x_0 \in \mathcal{R}^n$ and $w \in \mathcal{L}_2(\mathcal{R}_+)$. This means that there are constants, $\alpha, \epsilon > 0$, that are independent of T subject to $||P_Tz_1|| < \alpha||x_0||$ and subject to $||P_Tz_2||^2 - ||P_Tw||^2 \leq -\epsilon||P_Tw||^2$. Combining this with (37), we get

$$\begin{aligned} \Rightarrow ||P_Tz||^2 - ||P_Tw||^2 &\leq \alpha^2||x_0||^2 + (||P_Tz_2||^2 - ||P_Tw||^2) + 2\alpha||P_Tw||\,||x_0|| \\ &\leq \alpha^2||x_0||^2 - ||P_Tw||(\epsilon||P_Tw|| - 2\alpha||x_0||) \qquad (38) \\ &\leq (1 + \frac{1}{\epsilon})\alpha^2||x_0||^2 =: \beta||x_0||^2. \qquad (39) \end{aligned}$$

Here we have used the fact that (38) is maximized by $||P_Tw|| = \alpha||x_0||/\epsilon$. It is important to note that the right side of (39) does not depend on T. Therefore pick any terminal time T and set $w(t) = w_T^\star(x)(t)$. Clearly, this w is in $\mathcal{L}_2[0,T]$. From (32) we have that

$$x_0'X_T(0)x_0 \leq J_T(w_T^\star(x), u) \tag{40}$$

for every $u \in \mathcal{L}_2[0,T]$. In particular, let $u = Ky$; from (39) and (40) it follows that

$$\begin{aligned} &x_0'X_T(0)x_0 \leq \beta||x_0||^2 \\ &\Rightarrow X_T(0) \leq \beta I. \end{aligned} \tag{41}$$

As argued before, the game can be played starting at any time t and nothing will change. Indeed β is independent of both starting times and terminal times. Consequently,

$$X_T(t) \leq \beta I \quad \forall\ T \in \mathcal{R}_+ \quad \forall t \leq T. \tag{42}$$

It follows that $\{X_T(\cdot)\}$, indexed by T, is a nondecreasing sequence of continuous functions that is bounded above. Hence there exists a unique, bounded function $X_\infty(\cdot)$ subject to $\lim_{T\to\infty} X_T(t) = X_\infty(t)$ for every $t \in \mathcal{R}_+$. Note, in addition, that for each T the function $X_T(\cdot)$ is the solution to the finite time Riccati equation (30) and is continuous with respect to initial conditions. It follows (using the same arguments as in [12]), that $X_\infty(\cdot)$ satisfies the infinite horizon Riccati equation (16).

Stability of $A - (B_2B_2' - B_1B_1')X_\infty$. We will now show that X_∞ is stabilizing; i.e., the system Σ defined as $\dot{x}(t) = (A - (B_2B_2' - B_1B_1')X_\infty)(t)x(t)$ is exponentially stable. We will show later that there exist $\xi < \infty$ such that given any initial time t_0 and initial condition $x(t_0) = x_0$ we have

$$||(I - P_{t_0})B_1'X_\infty x|| \leq \xi||x_0|| \quad \forall t_0 \in \mathcal{R}_+. \tag{43}$$

Assuming that this is done, the stability of Σ follows from the following argument. Note that Σ can be written as $\dot{x}(t) = (A - B_2B_2'X_\infty)(t)x(t) + B_1B_1'X_\infty(t)x(t)$ with initial condition $x(t_0) = x_0$. Let Σ_{aux} be an auxiliary system defined as

$$\dot{x} = (A - B_2B_2'X_\infty)x + u.$$

As Σ_{aux} is exponentially stable, the corresponding input-output system, assuming x to be the output, is input-output stable (see [1]). This implies that there exist $\zeta < \infty$ subject to

$$||(I - P_{t_0})x|| \leq \zeta||(I - P_{t_0})u|| \quad \forall t_0 \in \mathcal{R}_+. \tag{44}$$

Using the fact that $u = B_1'X_\infty x$ and taking into account (43) and (44), we get that for all initial states $x(t_0) = x_0$,

$$||(I - P_{t_0})x|| \leq \xi\zeta||x_0||,$$

where the bound is independent of the initial time or state. The exponential stability of Σ follows from Theorem 3 of [4, p. 190].

We will now demonstrate the existence of a uniform bound ξ satisfying (43). To this end, recall the definitions of α and ϵ from (38). Set $\eta \geq (\alpha/\epsilon)(1 + \sqrt{1+\epsilon})$, and consider the system Σ. Because $X_\infty(t) := \lim_{T\to\infty} X_T(t)$, it is clear that for T large enough, the state trajectory of the system Σ_T defined as $\dot{x}_T(t) = (A - (B_2B_2' - B_1B_1')X_T)(t)x_T(t)$ will approach x in the $\mathcal{L}_2$ sense on the interval $[t_0, t_1]$ for some $t_1 < T$. At this point we assume, in contradiction to what we want to show, that there does not exist any bound satisfying (43). Therefore, given any $\eta > 0$ (in particular, the one chosen above) there exist x_0, t_0, T, and an interval $[t_0, t_1]$ with $t_1 < T$ such that $||B_1'X_Tx_T||_{[t_0,t_1]} > \eta||x_0||$ where the subscript on the norm has the obvious meaning. As $T > t_1$, clearly $||B_1'X_Tx_T||_{[t_0,T]} > \eta||x_0||$ also.

Now consider, over the interval $[t_0, T]$, system (31). In (38) replace u by $u^\star := -B_2'X_Tx_T$ and the w by $w^\star := B_1'X_Tx_T$ and denote the resulting output by $z^\star$. We have

$$||z^\star||^2_{[t_0,T]} - ||w^\star||^2_{[t_0,T]} \leq \alpha^2||x_0||^2 - ||w^\star||_{[t_0,T]}(\epsilon||w^\star||_{[t_0,T]} - 2\alpha||x_0||). \tag{45}$$

Using (32) it follows that

$$x_0'X_T(t_0)x_0 \leq \alpha^2||x_0||^2 - ||w^\star||_{[t_0,T]}(\epsilon||w^\star||_{[t_0,T]} - 2\alpha||x_0||). \tag{46}$$

Because we have chosen T to force $||w^\star||_{[t_0,T]} = ||B_1'X_Tx_T||_{[t_0,T]} > \eta$ it follows that

$$x_0'X_T(t_0)x_0 \leq \left(\alpha^2 - \frac{\alpha^2}{\epsilon}(1 + \sqrt{1+\epsilon})(1 + \sqrt{1+\epsilon}) - 2)\right)||x_0||^2 \leq 0. \tag{47}$$

This violates the nonnegativity of X_T, however. Hence we have a contradiction, thereby proving the existence of ξ satisfying (43). It now follows that X_∞ is indeed a stabilizing solution to the first Riccati equation (16).

Existence of Y_{tmp}. As the admissible controller K is linear and causal, clearly $P_T y = 0 \Rightarrow P_T u = 0$, for all $T \in \mathcal{R}_+$. Let $D_\perp(t)$ be a matrix function defined subject to $\binom{D_{21}(t)}{D_\perp(t)}$, is square, and

$$\left(\begin{array}{c} D_{21}(t) \\ D_\perp(t) \end{array}\right)\left(\begin{array}{c} D_{21}(t) \\ D_\perp(t) \end{array}\right)' = I. \tag{48}$$

Now set

$$w(t) := -D_{21}'(t)C_2(t)x(t) + B_1'(t)X_\infty(t)x(t) + D_\perp'(t)v_1(t) \quad \forall t \in [0, T), \tag{49}$$

where $v_1 \in \mathcal{L}_2[0, T]$ is arbitrary. It can be checked (using assumption (A2)) that with w chosen as above, $y(t) = 0$, for all $t \in [0, T]$, and system (7) becomes

$$\begin{array}{rcl} \dot{x}(t) & = & A_{tmp}(t)x(t) + B_1(t)D_\perp'(t)v_1(t), \\ y(t) & = & 0 \quad \forall t \in [0, T), \end{array} \tag{50}$$

where A_{tmp} is defined in (18). Recall that $v = u + B_2'X_\infty x$ and $r = w - B_1'X_\infty x$ for $t \in [0, T)$. These become $v = B_2'X_\infty x$ and $r = -D_{21}'C_2 x + D_\perp' v_1$, respectively. Integrating both sides of (22) from 0 to T gives us

$$\begin{aligned} ||P_T z||^2 - ||P_T w||^2 &= ||P_T v||^2 - ||P_T r||^2 - x'(T)X_\infty x(T) \\ &= ||P_T B_2'X_\infty x||^2 - ||P_T v_1||^2 - ||P_T C_2 x||^2 - x'(T)X_\infty x(T) \end{aligned} \tag{51}$$

because $x(0) = 0$. Consider once again the system over $[T, \infty)$ and with initial state $x(T)$. Set $w(t) := B_1'(t)X_\infty(t)\Phi_{A-(B_2B_2'+B_1B_1')X_\infty}(t, T)x(T)$ for $t \in [T, \infty)$. Since the system $\dot{x}(t) = (A - (B_2B_2' - B_1B_1')X_\infty)(t)x(t)$ is exponentially stable, $(I - P_T)w \in \mathcal{L}_2$. With this choice of w, the unique input that minimizes the cost $J(u, w) = ||(I - P_T)z||^2 - ||(I - P_T)w||^2$ for this *linear quadratic optimal control problem* is given by $u(t) = -B_2'(t)X_\infty(t)\Phi_{A-(B_2B_2'+B_1B_1')X_\infty}(t, T)x(T)$ for $t \in [T, \infty)$. Moreover, the optimal cost is given by $x'(T)X_\infty(T)x(T)$. This can be shown by noting that for any $(I - P_T)w \in \mathcal{L}_2$, the optimal u is obtained from the following two-point boundary value problem:

$$\left(\begin{array}{c} \dot{p} \\ \dot{x} \end{array}\right) = \left(\begin{array}{cc} -A' & C_1'C_1 \\ B_2B_2' & A \end{array}\right)\left(\begin{array}{c} p \\ x \end{array}\right) + \left(\begin{array}{c} 0 \\ B_1 w \end{array}\right),$$

$$x(T) = x(T), \qquad \lim_{t\to\infty} p(t) = 0,$$

$$u(t) = B_2'p(t).$$

The above claim now follows by noting that $p(t) = -X_\infty(t)\Phi_{A-(B_2B_2'+B_1B_1')X_\infty}(t, T)x(T)$ solves the above two-point boundary value problem.

Thus for any admissible controller K, with $u = Ky$ and with w chosen as above, we have

$$||(I - P_T)z||^2 - ||(I - P_T)w||^2 \geq x'(T)X_\infty(T)x(T). \tag{52}$$

By adding together (51) and (52) and by defining w over $\mathcal{R}_+$ as

$$w(t) := \begin{cases} -D_{21}'(t)C_2(t)x(t) + B_1'(t)X_\infty(t)x(t) + D_\perp'(t)v_1(t), & t < T, \\ B_1'(t)X_\infty(t)\Phi_{A-(B_2B_2'+B_1B_1')X_\infty}(t, T)x(T), & t \geq T, \end{cases}$$

we get, for the system in (50),

$$||z||^2 - ||w||^2 \geq ||P_T B_2' X_\infty x||^2 - ||P_T v_1||^2 - ||P_T C_2 x||^2. \tag{53}$$

(We note that this idea of splitting the input has been used in a similar context in [13].) But with $u = Ky$ we have $||T_{zw}|| < 1$ or $||z||^2 - ||w||^2 < 0$, for all $w \neq 0$. Hence it follows from (53) that

$$||P_T B_2' X_\infty x||^2 - ||P_T v_1||^2 - ||P_T C_2 x||^2 < 0 \quad \text{whenever } P_T v_1 \neq 0 \tag{54}$$

for system (50). We will now show that (54) is a sufficient condition for the existence of a solution to the second Riccati equation (17), which we reproduce here for ease of reference as follows:

$$\begin{aligned} \dot{Y}_{tmp}(t) &= A_{tmp}(t)Y_{tmp}(t) + Y_{tmp}(t)A_{tmp}'(t) - Y_{tmp}(t)(C_2'(t)C_2(t) \\ &\quad -X_\infty(t)B_2(t)B_2'(t)X_\infty(t))Y_{tmp}(t) + B_1(t)B_1'(t), \quad Y_{tmp}(0) = 0. \end{aligned}$$

Define the co-state p as follows:

$$\dot{p}(t) = -A_{tmp}'(t)p(t) + (C_2'C_2 - X_\infty B_2 B_2' X_\infty)(t)x(t), \qquad p(0) = p_0. \tag{55}$$

If we set $v_1 := D_\perp B_1' p$ then from (50) and (55) we get

$$\begin{pmatrix} \dot{p} \\ \dot{x} \end{pmatrix} = \begin{pmatrix} -A_{tmp}' & C_2'C_2 - X_\infty B_2 B_2' X_\infty \\ B_1 B_1' & A_{tmp} \end{pmatrix} \begin{pmatrix} p \\ x \end{pmatrix}, \quad \begin{pmatrix} p(0) \\ x(0) \end{pmatrix} = \begin{pmatrix} p_0 \\ 0 \end{pmatrix}. \tag{56}$$

Let the transition matrix of this system be given by

$$\begin{pmatrix} p(t) \\ x(t) \end{pmatrix} = \begin{pmatrix} \Phi_{11}(t,\tau) & \Phi_{12}(t,\tau) \\ \Phi_{21}(t,\tau) & \Phi_{22}(t,\tau) \end{pmatrix} \begin{pmatrix} p(\tau) \\ x(\tau) \end{pmatrix}. \tag{57}$$

We will now show that $\Phi_{11}(t,0)$ is nonsingular for all $t \in \mathcal{R}_+$. Suppose, on the contrary, that there is a time T and a vector $p_0 \neq 0$ subject to $p(T) = \Phi_{11}(T,0)p_0 = 0$. Choosing $v_1 = D_\perp B_1' p$ as before, two cases arise: (1) $P_T v_1 = 0$ and (2) $P_T v_1 \neq 0$. The first case is taken care of easily because if $P_T v_1 = 0$ then $P_T x = 0$, which implies that $\dot{p}(t) = -A_{tmp}'(t)p(t)$ and $p(T) = 0$ (see (56)). This clearly is impossible in a homogeneous system with a nonzero initial condition; hence we are led to the required contradiction. On the other hand, if $P_T v_1 \neq 0$ then we can differentiate the product $x'(t)p(t)$ along the trajectory of (56) and rearrange the resulting equation to get

$$\frac{d(x'p)}{dt} = v_1'v_1 + x'C_2'C_2 x - x'X_\infty B_2 B_2' X_\infty x. \tag{58}$$

Note that $x(0) = 0$ and that $p(T) = 0$; therefore by integrating (58) from 0 to T, we get

$$0 = ||P_T B_2' X_\infty x||^2 - ||P_T v_1||^2 - ||P_T C_2 x||^2, \tag{59}$$

which is a clear contradiction to (54). This proves that $\Phi_{11}(t,0)$ is nonsingular for all $t \in \mathcal{R}_+$.

We set $Y_{tmp}(t) := \Phi_{21}(t,0)\Phi_{11}^{-1}(t,0)$ and a simple calculation shows us that Y_{tmp} satisfies the second Riccati equation (17). All that remains to be shown is that Y_{tmp} has the required properties; i.e., it is nonnegative definite, bounded, and stabilizing.

Nonnegativity is proved by relating Y_{tmp} to M, the solution to the following filter Riccati equation:

$$\dot{M}(t) = A_{tmp}(t)M(t) + M(t)A'_{tmp}(t) - M(t)C'_2(t)C_2(t)M(t) + B_1(t)B'_1(t), \quad M(0) = 0. \tag{60}$$

It can be shown that (A_{tmp}, C_2) is detectable.[1] This is sufficient to ensure that the Riccati equation above has a nonnegative, bounded solution. Clearly, $Y_{tmp}(t) \geq M(t)$, for all $t \in \mathcal{R}_+$ and because $M(t) \geq 0$ it follows that Y_{tmp} is nonnegative.

Boundedness of Y_{tmp}. Next we show that there is a number $\nu < \infty$ subject to $Y_{tmp}(t) \leq \nu I$, for all $t \in \mathcal{R}_+$. Note that for the entire discussion that follows we have assumed that $w(t)$ is chosen as in (49), i.e., to make $P_T y = 0$. Assume that $x(T)$ is a reachable state for system (50) at time $t = T$. Then, from Lemma 2.3, we have

$$\inf_{||P_T v_1|| < \infty} (||P_T C_2 x||^2 + ||P_T v_1||^2, \quad x(0) = 0, \quad x(T) = x(T)) \geq \zeta ||x(T)||^2, \tag{61}$$

where $\zeta > 0$ is independent of T. Let T be any terminal time subject to $Y_{tmp}(T) \neq 0$. Moreover, assume as before that the input to (50) is given by $v_1 = D_\perp B'_1 p$, where p satisfies the differential equation in (55). It is easy to see that the following hold up to $t = T$:

$$r(t) = B'_1(t)p(t) - D'_{21}(t)C_2(t)x(t), \tag{62}$$

$$x(t) = Y_{tmp}(t)p(t). \tag{63}$$

This means that $x(T)$ is reachable at $t = T$ using this choice of v_1 if and only if it is in the range space of $Y_{tmp}(T)$. Calculate the derivative of the quantity $x'(t)p(t)$ along the trajectory of (56), and integrate from 0 to T to get

$$x'(T)p(T) = ||P_T r||^2 - ||P_T v||^2. \tag{64}$$

From Lemma 2.2 it follows that with $u = Ky$, $||T_{zw}|| < 1$ implies $||T_{vr}|| < 1$.[2] This means that there exists $\varepsilon > 0$ subject to

$$||P_T v||^2 - ||P_T r||^2 \leq -\varepsilon ||P_T r||^2 \quad \forall ||P_T r|| \neq 0. \tag{65}$$

Clearly, (64) and (65) imply that $x'(T)p(T) \geq \varepsilon ||P_T r||^2$. Coupling this with (61), we get

$$x'(T)p(T) \geq \varepsilon \zeta x'(T)x(T). \tag{66}$$

Setting $x(T) := \lambda e$, where e is the eigenvector corresponding to the maximum eigenvalue of $Y_{tmp}(T)$ and λ is a scaling factor, we can make $x'(T)p(T) = (1/||Y_{tmp}(T)||)x'(T)x(T)$. Now by defining $\nu := 1/(\varepsilon\zeta)$ we have from the above equation

$$||Y_{tmp}(t)|| \leq \nu I \quad \forall t \in \mathcal{R}_+. \tag{67}$$

Stability of $A_{tmp} - Y_{tmp}(C'_2 C_2 - X_\infty B_2 B'_2 X_\infty)$. Finally, we prove that Y_{tmp} is stabilizing. From (56) and (63) with v_1 chosen as before ($v_1 = D_\perp B'_1 p$), we have

$$\dot{p}(t) = (-A'_{tmp} + (C'_2 C_2 - X_\infty B_2 B'_2 X_\infty)Y_{tmp})(t)p(t). \tag{68}$$

[1] See the Appendix.

[2] See the Appendix.

Let the boundary condition for (68) be $p(T) = p_T$. Clearly, given *any* $p_T \in \mathcal{R}^n$, there exists an initial condition p_0 in (56) subject to $p(T) = p_T$. Now existence of an admissible controller as in (65) implies

$$(69) \qquad ||P_T v||^2 - ||P_T r||^2 \leq -\varepsilon||P_T r||^2 \leq -\varepsilon||P_T v||^2 \quad \forall ||P_T r|| \neq 0.$$

Noting that $v(t) = B_2' X_\infty x(t) = B_2' X_\infty Y_{tmp} p(t)$ and r is given by (62), the above equation becomes

$$(70) \quad ||P_T B_2' X_\infty Y_{tmp} p||^2 \leq \frac{1}{\varepsilon}(||P_T B_1' p||^2 + ||P_T C_2 Y_{tmp} p||^2 - ||P_T B_2' X_\infty Y_{tmp} p||^2)$$

for system (68). Now, by setting $t^* = -t$, we get a time-reversed system for (68) defined as follows:

$$(71) \quad \begin{array}{rcl} \dot{p}(t^*) &=& (A_{tmp}' - (C_2'C_2 - X_\infty B_2 B_2' X_\infty)Y_{tmp})(t^*)p(t^*), \qquad p(-T) = p_T, \\ v(t^*) &=& B_2'(t^*)X_\infty(t^*)Y_{tmp}(t^*)p(t^*), \end{array}$$

where p_T is arbitrary. A repetition of the ideas in the sufficiency part, suitably dualized, shows that $\dot{x}(t^*) = (A_{tmp}' - C_2'C_2Y_{tmp})(t^*)x(t^*)$ is stable on $\mathcal{R}_-$. This means that the system defined in (71) is detectable. Observe that for the reversed time dual system given by (71), the equivalent of (69) and of (70) are

$$(72) \quad \begin{array}{rl} ||v||_{[-T,0]}^2 - ||r||_{[-T,0]}^2 &\leq -\varepsilon||r||_{[-T,0]}^2 \\ &\leq -\varepsilon||v||_{[-T,0]}^2 \end{array}$$

and

(73)

$$||B_2' X_\infty Y_{tmp} p||_{[-T,0]}^2 \leq \frac{1}{\varepsilon}(||B_1' p||_{[-T,0]}^2 + ||C_2 Y_{tmp} p||_{[-T,0]}^2 - ||B_2' X_\infty Y_{tmp} p||_{[-T,0]}^2),$$

respectively. Next consider the quantity $p'(t^*)Y_{tmp}(t^*)p(t^*)$ and calculate its derivative along the trajectory of (71). Substituting for $\dot{Y}_{tmp}(t^*)$ from (25) and integrating the resulting expression from $-T$ to 0, we get

$$(74) \quad ||B_1' p||_{[-T,0]}^2 + ||C_2 Y_{tmp} p||_{[-T,0]}^2 - ||B_2' X_\infty Y_{tmp}||_{[-T,0]}^2 = p_T' Y_{tmp}(-T) p_T.$$

Combining equations (67), (73), and (74) gives us

$$(75) \qquad ||B_2' X_\infty Y_{tmp} p||_{[-T,0]}^2 \leq \frac{\nu}{\varepsilon}||p_T||^2.$$

Note that ν/ε is independent of T. Equation (75) then simply states that the output of (71) is *uniformly* bounded with respect to starting times and initial conditions. As the system is detectable, this translates to uniform bound on the state of the system itself. In other words, using the fact that $(A_{tmp}' - C_2'C_2Y_{tmp})$ is exponentially stable and that system (71) can be equivalently written as

$$\begin{array}{rcl} \dot{p}(t^*) &=& (A_{tmp}' - (C_2'C_2 Y_{tmp})p(t^*) + X_\infty B_2 B_2' X_\infty)Y_{tmp})(t^*)p(t^*), \qquad p(-T) = p_T, \\ v(t^*) &=& B_2'(t^*)X_\infty(t^*)Y_{tmp}(t^*)p(t^*), \end{array}$$

we can easily derive a bound ζ, independent of T subject to

$$(76) \qquad ||p||_{[-T,0]}^2 \leq \zeta||p_T||^2.$$

By appropriately dualizing the arguments found in the proof of Theorem 3 of [4, p. 190], we can conclude from (76) that

$$\dot{p}(t^*) = (A'_{tmp} - (C'_2C_2 - X_\infty B_2B'_2X_\infty)Y_{tmp})(t^*)p(t^*)$$

is stable on $\mathcal{R}_-$ (the interested reader may also see [16]). Now we are done, as it follows that the dual of this system, defined, as usual, as

$$\dot{p}^*(t) = (A_{tmp} - Y_{tmp}(C'_2C_2 - X_\infty B_2B'_2X_\infty))(t)p^*(t),$$

is also exponentially stable. □

Appendix. In this Appendix we establish two important facts that we have used in the latter half of the necessity part of the proof but have not yet proved. Recall that in the first half of the necessity part we proved that the existence of an admissible controller implies that there is a bounded stabilizing solution X_∞ to the first Riccati equation. Given such an X_∞ we show, first, that the closed-loop operator T_{vr} is stable and $||T_{vr}|| < 1$, and second, that (A_{tmp}, C_2) is detectable. To begin with, note that

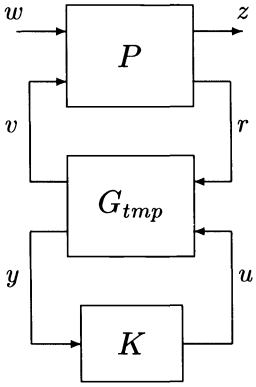


FIG. 4. *An equivalent representation.*

Fig. 1 can be redrawn as Fig. 4, where P is as defined in (20) and G_{tmp} is defined as follows:

$$\begin{aligned} \dot{x}_{tmp}(t) &= A_{tmp}(t)x_{tmp}(t) + B_1(t)r(t) + B_2(t)u(t), \\ v(t) &= B'_2(t)X_\infty(t)x_{tmp}(t) + u(t), \\ y(t) &= C_2(t)x_{tmp}(t) + D_{21}(t)r(t). \end{aligned} \tag{77}$$

Because K stabilizes the closed loop it can be shown that the modified system in Fig. 4 is also exponentially stable. The main idea is as follows. Let us denote by T_{vr} the closed-loop combination of G_{tmp} and K. Then, given any admissible linear controller K with a stabilizable and detectable realization, it is routine to verify that the realization for T_{vr} is also stabilizable and detectable. Similarly, the feedback connection of P and T_{vr} is verified to be stabilizable from w and detectable from z. Because T_{zw} is input-output stable and the corresponding realization is stabilizable and detectable, it follows that T_{zw} is (internally) exponentially stable as well.

Next, a direct calculation shows that P_{21} is stably invertible. Now it is clear that P and T_{vr} satisfy all the assumptions of Lemma 2.2 (with T_{vr} replacing Q). It follows that, as the closed loop is exponentially stable and $||T_{zw}|| < 1$, the system T_{vr} is stable and, furthermore, that $||T_{vr}|| < 1$. This establishes the first part.

For the second part note that because T_{vr} is internally exponentially stable, it follows that the system G_{tmp} is exponentially stabilized by K. It immediately follows that the realization given in (77) is stabilizable from u and detectable from y. In particular, (A_{tmp}, C_2) is detectable.

REFERENCES

[1] B. D. O. Anderson, *Internal and external stability of linear time-varying systems*, SIAM J. Control Optim., 20 (1982), pp. 408–413.

[2] B. D. O. Anderson and J. B. Moore, *Detectability and stabilizability of time-varying discrete-time linear systems*, SIAM J. Control Optim., 19 (1981), pp. 20–32.

[3] ———, *Optimal Control: Linear Quadratic Methods*, Prentice–Hall, Englewood Cliffs, NJ, 1990.

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

[5] J. C. Doyle, K. Glover, P. P. Khargonekar, and B. A. Francis, *State-space solutions to standard H_2 and H_∞ control problems*, IEEE Trans. Automat. Control, 34 (1989), pp. 831–847.

[6] A. Feintuch and B. A. Francis, *Uniformly optimal control of linear time-varying systems*, Systems Control Lett., 5 (1984), pp. 67–71.

[7] ———, *Uniformly optimal control of linear feedback systems*, Automatica, 21 (1985), pp. 563–574.

[8] A. Feintuch, P. P. Khargonekar, and A. Tannenbaum, *On the sensitivity minimization problem for linear time-varying periodic systems*, SIAM J. Control Optim., 24 (1986), pp. 1076- 1085.

[9] B. A. Francis, *A Course in H_∞ Control Theory*, Lecture Notes in Control and Inform. Sci., Vol. 88, Springer-Verlag, New York, 1987.

[10] T. T. Georgiou and P. P. Khargonekar, *A constructive algorithm for sensitivity optimization of periodic systems*, SIAM J. Control Optim., 25 (1987), pp. 334–340.

[11] K. Glover and J. C. Doyle, *State-space formulae for all stabilizing controllers that satisfy an H_∞-norm bound and relations to risk sensitivity*, Systems Control Lett., 11 (1988), pp. 167–172.

[12] R. E. Kalman, *Contributions to the theory of optimal control*, Bol. Soc. Mat. Mexicana (2), 5 (1960), pp. 102–119.

[13] P. P. Khargonekar, *State-space H_∞ control theory*, Tech. Report, Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, November 1989.

[14] P. P. Khargonekar, I. R. Petersen, and K. Zhou, *Robust stabilization of uncertain linear systems: quadratic stabilization and H^∞ control theory*, IEEE Trans. Automat. Control, 35 (1990), pp. 356–361.

[15] D. J. N. Limebeer, B. D. O. Anderson, P. P. Khargonekar, and M. Green, *A game theoretic approach to H^∞ control for time varying systems*, Tech. Report, Department of Electrical Engineering, Imperial College, London, 1989.

[16] R. Ravi, A. M. Pascoal, and P. P. Khargonekar, *Normalized coprime factorizations and the graph metric for linear time-varying systems*, in Proc. 29th Conference on Decision and Control, Honolulu, HI, December 1990, pp. 1241–1246.

[17] M. G. Safonov, D. J. N. Limebeer, and R. Y. Chiang, *Simplifying the H_∞ theory via loop shifting, matrix pencil, and descriptor concepts*, Internat. J. Control, 50 (1989), pp. 2467–2488.

[18] C. Scherer, *H_∞-optimization without assumptions on finite or infinite zeros*, Tech. Report, Mathematisches Institut Am Hubland, October 1989. SIAM J. Control Optim., 30 (1992), to appear.

[19] A. A. Stoorvogel and H. Trentelman, *The quadratic matrix inequality and in singular H_∞ control with state feedback*, SIAM J. Control Optim., 28 (1990), pp. 1190–1208.

[20] G. Tadmor, *The standard H_∞ problem and the maximum principle: The general linear case*, Tech. Report 192, University of Texas at Dallas, May 1989.

[21] ———, *Worst-case design in the time domain: The maximum principle and the standard H_∞ problem*, Math. Control Signals Systems, 3 (1989), pp. 301–324.

[22] S. Takeda and A. R. Bergen, *Instability of feedback systems by orthogonal decomposition of* L_2, IEEE Trans. Automat. Control, 18 (1973), pp. 631–636.

[23] L. Y. Wang and G. Zames, *Local-global double algebras for slow* H^∞ *adaptation*, in Proc. 28th Conference on Decision and Control, Tampa, FL, December 1989.

[24] G. Zames, *Feedback and optimal sensitivity: model reference transformations, multiplicative seminorms, and approximate inverses*, IEEE Trans. Automat. Control, 26 (1981), pp. 301–320.

[25] K. Zhou and P. P. Khargonekar, *An algebraic Riccati equation approach to* H_∞ *optimization*, Systems Control Lett., 11 (1988), pp. 85–92.

SIAM CONTROL AND OPTIMIZATION
Vol. 29, No. 6, pp. 1414–1442, November 1991

008

VELOCITY METHOD AND LAGRANGIAN FORMULATION FOR THE COMPUTATION OF THE SHAPE HESSIAN*

MICHEL C. DELFOUR† AND JEAN-PAUL ZOLÉSIO‡

Abstract. The object of this paper is to study the shape Hessian of a shape functional by the velocity (speed) method. It contains a review and an extension of the velocity method and its connections with methods using first- or second-order perturbations of the identity. The key point is that all these methods yield the same shape gradient but different and unequal shape Hessian since each method depends on a choice of "connection." However, for autonomous velocity fields the velocity method yields a canonical bilinear Hessian. Expressions obtained by other methods can be recovered by adding to that canonical term the shape gradient acting on the acceleration of the velocity field associated with the choice of perturbation of the identity. The second part of the paper is an application of the Lagrangian method with function space embedding to compute the shape gradient and Hessian of a simple cost function associated with the nonhomogeneous Dirichlet problem.

Key words. shape optimization, velocity method, Hessian, second-order derivatives

AMS(MOS) subject classifications. 49A22

1. Introduction.

The object of this paper is to study the *shape Hessian* by the *velocity* (*speed*) *method* (cf. Céa [1]–[3] and Zolésio [1], [2]) and to apply the *Lagrangian method* with *function space embedding* to compute the shape gradient and Hessian of a simple cost function associated with the nonhomogeneous Dirichlet problem. To do this we introduce a general method that applies to *differentiable semiconvex* cost functionals with applications to more general problems than the simple illustrative example we have chosen to consider. We emphasize the use of the *function space embedding* method (cf. Delfour and Zolésio [1]–[4], [7]) combined with the implicit use of Lagrange multipliers. This paper complements our previous work, where we have used a variational formulation and *function space parametrization* for the Neumann problem (cf. Zolésio and Delfour [8]).

In shape sensitivity analysis the size of the computations can quickly become quite large. Therefore, it is extremely important to know and understand the fundamental structure of the shape gradient and the shape hessian to simplify the computations and obtain mathematically meaningful expressions. For this reason we systematically revise and update the velocity (speed) method and then present new results for the second-order shape derivative. In the process we show how to associate with methods of *perturbation of the identity* (first and second order) an appropriate nonautonomous family of velocity fields. For the shape gradient, the different methods yield expressions that may look different but are all equal. However, this is no longer true for the

*Received by the editors February 12, 1990; accepted for publication (in revised form) September 28, 1990.

†Centre de Recherches Mathématiques et Département de Mathématiques et de Statistique, Université de Montréal, CP 6128–A, Montréal (QC) H3C 3J7, Canada. The research of this author was supported in part by a Killam Fellowship from the Canada Council, Natural Sciences and Engineering Research Council of Canada grants OGP–8730 and INF–7939 and a FCAR grant (Fonds pour la Formation de Chercheurs et l'Aide à la Recherche, Québec) from the "Ministère de l'Education du Québec."

‡Institut Non linéare de Nice, Faculté des sciences, Parc Valrose, F-06034 Nice-Cédex, France.

shape Hessian. In fact, we will show in §2.4 that different perturbations of the identity can yield different final expressions that are not equal. The potentially confusing consequence of this fact is that we can introduce an infinity of definitions based on perturbations of the identity. However, we will show that they always contain a *canonical bilinear term* plus the shape gradient of the functional acting in the direction of an *acceleration* field, which is characteristic of the chosen perturbation. The canonical bilinear term exactly coincides with the second-order shape derivative obtained by the velocity (speed) method for autonomous velocity fields. Each expression arising from a perturbation of the identity can be strictly recovered by adding to the canonical term the shape gradient acting in the direction of an appropriate acceleration field. Therefore, we propose to refer to this canonical term as the *shape Hessian*.

The above considerations clarify the fundamental concepts and reduce their complexity, but they do not eliminate all the associated computations. We need methods that provide both quick formal computations and appropriate mathematical justifications. We use Lagrangian methods combined with the use of theorems on the derivative of a MinMax with respect to a parameter. Such methods are well known and extensively used in mechanical sciences, mathematical programming, and optimal control theory. Their application to shape sensitivity analysis is not completely straightforward since it leads to the *time-dependence* of the underlying function spaces appearing in the MinMax formulation. This phenomenon seems to be specific to that class of problems. Two techniques are available to get around this difficulty: the *function space parametrization* and the *function space embedding* methods. The first one has been used in Delfour and Zolésio [8], [9], the second one will be used here.

It is fair to say that the use of shape Hessians for discretized finite element models and finitely parametrized shapes have been used in many places in the engineering and mechanics literature. Some numerical expertise is available (cf., for instance, Bern [1], Bern, Chenot, Demay, and Zolésio [1]) and it is suspected that the really performing algorithms are not available in the open literature since they are marketable industrial products.

A few papers have dealt with the second variation of a shape cost function for linear partial differential equations models. To our knowledge, the first one by Fujii [1] used a second-order perturbation of the identity along the normal to the boundary for second-order linear elliptic problems. An extremely interesting paper by Arumugan and Pironneau [1], [2] used the shape second variation to solve the "ribblet problem." Finally, Simon [1] presented a computation of the second variation using a first-order perturbation of the identity. The first general approach to the computation of shape Hessians can be found in Delfour and Zolésio [8], [9]. It uses the velocity (speed) method and includes a simple illustrative example for the Neumann problem.

In conclusion, we would like to insist that the velocity method and methods using first- and second-order perturbations of the identity lead to three different second-order shape derivatives that are not equal. The velocity method with autonomous velocity fields provides the canonical bilinear shape Hessian and all the other derivatives can be recovered by special choices of nonautonomous velocity fields.

2. Shape derivative: definitions and properties.

In this section we recall and extend the definitions of a shape gradient and a shape Hessian based on the velocity method (cf. Zolésio [1], [2], Delfour and Zolésio [8], [9]) and discuss their relationship to various methods based on perturbations of the identity operator. The proof of each theorem is given in the Appendix.

2.1. Velocity (speed) method and perturbations of the identity operator. Let $V : [0,\tau] \times \mathbb{R}^N \to \mathbb{R}^N$ be a given velocity field for some fixed $\tau > 0$. The map V can be viewed as a family $\{V(t)\}$ of nonautonomous velocity fields on $\mathbb{R}^N$ defined by

$$x \mapsto V(t)(x) \overset{\text{def}}{=} V(t,x) : \mathbb{R}^N \mapsto \mathbb{R}^N. \tag{1}$$

Assume that

$$\text{(V)} \qquad \begin{cases} \forall x \in \mathbb{R}^N, & V(\cdot,x) \in C^0([0,\tau];\mathbb{R}^N), \\ \exists c > 0, \quad \forall x,y \in \mathbb{R}^N, & \|V(\cdot,y) - V(\cdot,x)\|_{C^0([0,\tau];\mathbb{R}^N)} \le c|y-x|, \end{cases}$$

where $V(\cdot,x)$ denotes the function $t \mapsto V(t,x)$. Associate with V the solution $x(t;X)$ of the ordinary differential equation

$$\frac{dx}{dt}(t) = V(t,x(t)), \quad t \in [0,\tau], \quad x(0) = X \in \mathbb{R}^N \tag{2}$$

and introduce the family of homeomorphisms

$$X \mapsto T_t(V)(X) \overset{\text{def}}{=} x(t,X) : \mathbb{R}^N \to \mathbb{R}^N \tag{3}$$

and the maps

$$(t,X) \mapsto T_V(t,X) \overset{\text{def}}{=} T_t(V)(X) : [0,\tau] \times \mathbb{R}^N \to \mathbb{R}^N, \tag{4}$$

$$(t,x) \mapsto T_V^{-1}(t,x) \overset{\text{def}}{=} T_t^{-1}(V)(x) : [0,\tau] \times \mathbb{R}^N \to \mathbb{R}^N. \tag{5}$$

By definition $T_0(X) = X$ and $T_0 = I$. Moreover, T_t is an evolution operator that verifies the usual semigroup property.

Note 2.1. In the sequel we will drop the V in $T_V(t,X)$ and $T_t(V)$ whenever no confusion is possible.

THEOREM 2.1. (i) *Under assumptions* (V) *on the map* V, *the maps* T *defined by* (3) *and* (4) *have the following properties*:

(T1)

$$\begin{aligned} &\forall X \in \mathbb{R}^N, \quad T(\cdot,X) \in C^1([0,\tau];\mathbb{R}^N), \\ &\exists c > 0, \quad \forall X,Y \in \mathbb{R}^N, \quad \|T(\cdot,Y) - T(\cdot,X)\|_{C^1([0,\tau];\mathbb{R}^N)} \le c|Y-X|, \end{aligned}$$

(T2)

$$\forall t \in [0,\tau], \quad X \mapsto T_t(X) = T(t,X) : \mathbb{R}^N \to \mathbb{R}^N \quad \textit{is bijective},$$

(T3)

$$\begin{aligned} &\forall x \in \mathbb{R}^N, \quad T^{-1}(\cdot,x) \in C^0([0,\tau];\mathbb{R}^N) \\ &\exists c > 0, \quad \forall x,y \in \mathbb{R}^N, \quad \|T^{-1}(\cdot,y) - T^{-1}(\cdot,x)\|_{C^0([0,\tau];\mathbb{R}^N)} \le c|y-x|. \end{aligned}$$

(ii) *If there exists a real number* $\tau > 0$ *and a map* $T : [0,\tau] \times \mathbb{R}^N \to \mathbb{R}^N$ *verifying assumptions* (T1), (T2), *and* (T3), *then the map*

$$(t,x) \mapsto V(t,x) = \frac{\partial T}{\partial t}(t, T_t^{-1}(x)) : [0,\tau] \times \mathbb{R}^N \to \mathbb{R}^N, \tag{6}$$

verifies assumptions (V), *where* T_t^{-1} *is the inverse of* $X \mapsto T_t(X)$.

This first theorem is an equivalence result. It says that we can either start from a family of velocity fields $\{V(t)\}$ on $\mathbb{R}^N$ or a family of transformations $\{T_t\}$ of $\mathbb{R}^N$ provided that the map $V, V(t,x) = V(t)(x)$, verifies (V) or the map $T, T(t,X) = T_t(X)$ verifies assumptions (T1), (T2), and (T3).

When we start from V, we obtain the velocity method. Given an initial domain Ω, the family of homeomorphisms $T_t(V)$ defines a family of transformed domains

$$\Omega_t = T_t(V)(\Omega) = \{T_t(V)(X) : X \in \Omega\}. \tag{7}$$

In examples where we start from T, it is usually possible to verify hypotheses (T1),(T2), and (T3) and construct the corresponding velocity field V defined in (6). For instance, perturbations of the identity to the first- or second-order fall in that category:

$$T(t,X) = X + tU(X) + \frac{t^2}{2}A(X) \quad (A = 0 \text{ for the first order}), \quad t \geq 0, \ X \in \mathbb{R}^N, \tag{8}$$

where U and A are given transformations of $\mathbb{R}^N$. It turns out that, for Lipschitz transformations U and A, we can construct a $\tau > 0$ for which hypotheses (T1), (T2), and (T3) are verified.

THEOREM 2.2. *Let U and A be two uniform Lipschitz transformations of* $\mathbb{R}^N$:

$$\exists c > 0, \quad \forall X,\ Y \in \mathbb{R}^N, \quad |U(Y) - U(X)| \leq c|Y - X|, \quad |A(Y) - A(X)| \leq c|Y - X|.$$

(i) *Let* $\tau = \min\{1, 1/4c\}$ *and T be given by* (8). *Then the velocity*

$$(t,x) \mapsto V(t,x) = U(T_t^{-1}(x)) + tA(T_t^{-1}(x)) : [0,\tau] \times \mathbb{R}^N \to \mathbb{R}^N, \tag{9}$$

on $[0,\tau]$ *verifies assumptions* (V).

Remark 2.1. This theorem only says that for Lipschitzian fields U and A there exists a $\tau > 0$, possibly very small, such that $\{T_t(X) :$ for all $t \in [0,\tau]\}$ is solution of the differential equation (2) for V given by (9). This is all we need to define a shape derivative and this is the general approach followed in the various papers using perturbations of the identity operator. So we do not need to restrict our hypotheses to autonomous linear and nilpotent U's and A's. Higher-order perturbations of the identity operator can be considered, and Theorem 2.2 will still apply for some sufficiently small $\tau > 0$. However, they are not necessary for first- and second-order derivatives.

Remark 2.2. Observe that from (8) and (9)

$$V(0) = U, \ \overset{\bullet}{V}(0)(x) \overset{\text{def}}{=} \frac{\partial V}{\partial t}(t,x)|_{t=0} = A - [DU]U, \tag{10}$$

where DU is the Jacobian matrix of U. The transformation $\overset{\bullet}{V}(0)$ of $\mathbb{R}^N$ is an *acceleration* field at $t = 0$, which will always be present even when $A = 0$.

2.2. Shape gradient. In general a, *shape functional* will be a map

$$\Omega \mapsto J(\Omega) : \mathcal{A} \subset \mathcal{P}(\mathbb{R}^N) \to \mathbb{R} \tag{11}$$

defined on a subset $\mathcal{A}$ of the set $\mathcal{P}(\mathbb{R}^N)$ of all subsets of $\mathbb{R}^N$. Under the action of a velocity V verifying assumptions (V), the domain Ω is transformed into a new domain $\Omega_t(V) = T_t(V)(\Omega)$.

DEFINITION 2.1. Given a velocity field V verifying assumptions (V), J is said to have an *Eulerian semiderivative* at Ω in the direction V if the following limit exists and is finite:

$$\lim_{t \searrow 0} [J(\Omega_t(V)) - J(\Omega)]/t. \tag{12}$$

Whenever it exists, it is denoted by $dJ(\Omega; V)$.

Remark 2.3. This is the original terminology introduced by Céa [1],[2] and Zolésio [1] in 1979. It is now widely recognized and used in all papers based on the velocity (speed) method.

This definition is quite general and covers situations where $dJ(\Omega; V)$ is a function of the whole family of velocity fields $\{V(t) : t \in [0, \tau]\}$. However, in most applications $dJ(\Omega; V)$ will only depend on $V(0)$, the velocity field at $t = 0$. This is a very important property since $dJ(\Omega; V)$ can then be obtained by using the autonomous vector field W,

$$W(t, x) = V(0, x) \quad \forall x \in \mathbb{R}^N \quad \forall t \geq 0,$$

instead of the nonautonomous field V. It is customary to use the notation $dJ(\Omega; V(0))$ for the Eulerian semiderivative computed for the autonomous field W and we make the identification $W \equiv V(0)$.

It turns out that the very important property $dJ(\Omega; V) = dJ(\Omega; V(0))$ can be obtained under a simple continuity hypothesis on the map $V \mapsto dJ(\Omega; V)$. This, of course, requires the introduction of a special topology on the space of velocity fields V. To be more precise, we introduce some notation. For any integers $k \geq 0$ and $m \geq 0$, and any compact subset K of $\mathbb{R}^N$

$$\mathcal{V}_K^{m,k} \;=\; C^m([0, \tau]; \mathcal{D}^k(K, \mathbb{R}^N)) \cap \mathcal{L}, \tag{13}$$

where $\mathcal{D}^k(K, \mathbb{R}^N)$ is the space of all k-times continuously differentiable maps from $\mathbb{R}^N$ to $\mathbb{R}^N$ with compact support in K and

$$\mathcal{L} = \{V : [0, \tau] \times \mathbb{R}^N \to \mathbb{R}^N : V \text{ verifies assumption (V)}\}. \tag{14}$$

When $k = \infty$ we drop the superscript ∞ and simply write $\mathcal{D}(K, \mathbb{R}^N)$ instead of $\mathcal{D}^\infty(K, \mathbb{R}^N)$. With the above definitions, we introduce the following new space:

$$\overset{\rightarrow}{\mathcal{V}}^{m,k} \overset{\text{def}}{=} \underset{K}{\underset{\rightarrow}{\lim}} \left\{\mathcal{V}_K^{m,k} : \forall K \;\text{ compact } \mathit{in}\ \mathbb{R}^N\right\}, \tag{15}$$

where $\underset{\rightarrow}{\lim}$ denotes the inductive limit endowed with its natural inductive limit topology. This is not a Fréchet space. For autonomous fields, the above constructions reduce to

$$\mathcal{V}^k = \mathcal{D}^k(\mathbb{R}^N, \mathbb{R}^N) \cap \text{ Lip}(\mathbb{R}^N, \mathbb{R}^N), \tag{16}$$

where Lip $(\mathbb{R}^N, \mathbb{R}^N)$ denotes the space of transformations of $\mathbb{R}^N$ that are uniformly Lipschitzian. Again, we will use the notation $\mathcal{D}(\mathbb{R}^N, \mathbb{R}^N)$ for $\mathcal{D}^\infty(\mathbb{R}^N, \mathbb{R}^N)$. In all cases, assumptions (V) will be verified.

THEOREM 2.3. *Let Ω be a domain in $\mathbb{R}^N$ and $m \geq 0$ and $k \geq 0$ be integers. Assume that for all V in $\overrightarrow{\mathcal{V}}^{m,k}$, $dJ(\Omega; V)$ exists and that the map*

$$V \mapsto dJ(\Omega; V) : \overset{\rightarrow}{\mathcal{V}}^{m,k} \to \mathbb{R} \tag{17}$$

is continuous. Then

$$\forall V \in \overset{\rightarrow}{\mathcal{V}}^{m,k}, \quad dJ(\Omega; V) = dJ(\Omega; V(0)). \tag{18}$$

In the above analysis we have chosen to follow the classical framework of the theory of distributions (cf. Schwartz [1]) and perturb the domain Ω by velocity fields V with compact support. This means that we simultaneously deal with bounded and unbounded domains Ω's. Theorem 2.3 is a generalization of the earlier result of Zolésio [1].

DEFINITION 2.2. Let Ω be a domain in $\mathbb{R}^N$.

(i) The shape functional J is said to be *shape differentiable* at Ω if the Eulerian semiderivative $dJ(\Omega; V)$ exists for all V in $\mathcal{D}(\mathbb{R}^N, \mathbb{R}^N)$ and the map

$$V \mapsto dJ(\Omega; V) : \mathcal{D}(\mathbb{R}^N, \mathbb{R}^N) \to \mathbb{R} \tag{19}$$

is linear and continuous.

(ii) The map (19) defines a vector distribution $G(\Omega)$, which will be called the *shape gradient* of J at Ω.

(iii) If there exists a finite $k \geq 0$ such that $G(\Omega)$ is continuous for the $\mathcal{D}^k(\mathbb{R}^N, \mathbb{R}^N)$-topology, we say that $G(\Omega)$ is of finite order qk.

The next theorem gives additional properties of shape differentiable functionals.

THEOREM 2.4 (Generalized Hadamard's structure theorem). *Let Ω be an open domain in $\mathbb{R}^N$ with boundary Γ and assume that J is shape differentiable.*

(i) *The support of $G(\Omega)$ is contained in Γ. Moreover, when the support of $G(\Omega)$ is compact the order of $G(\Omega)$ is finite.*

(ii) *If $G(\Omega)$ is of finite order $k \geq 0$ and Ω is an open domain in $\mathbb{R}^N$ with boundary Γ in C^{k+1}, then there exists a scalar distribution $g(\Omega)$ in $\mathcal{D}^k(\Gamma)'$ such that*

$$dJ(\Omega; V) = < g(\Omega),\ \gamma_\Gamma V \bullet n >_{\mathcal{D}^k(\Gamma)}, \tag{20}$$

where $\gamma_\Gamma V$ is the trace of V on Γ, n is the unit outward normal to Ω on Γ, and $V \bullet n$ denotes the scalar product of V and n in $\mathbb{R}^N$.

The name of the theorem comes from the famous prized paper by Hadamard [1], written in 1907, where he used velocity fields along the normal to the boundary Γ of a C^∞ domain to compute the derivative of the first eigenvalue of the plate. Theorem 2.4 was proved by Zolésio [1] in 1979. A new proof using Nagumo's [1] theorem is given in the Appendix.

Remark 2.4. When Γ is compact, $\mathcal{D}^k(\Gamma)$ coincides with $C^k(\Gamma)$.

2.3. Shape Hessian. We first study the second-order Eulerian semiderivative $d^2J(\Omega; V; W)$ of a functional $J(\Omega)$ for two nonautonomous vector fields V and W. A first theorem shows that under some natural continuity hypotheses, $d^2J(\Omega; V; W)$ is the sum of two terms: a "canonical term $d^2J(\Omega; V(0); W(0))$" plus the first-order Eulerian semiderivative $dJ(\Omega; \overset{\bullet}{V}(0))$ at Ω in the direction $\overset{\bullet}{V}(0)$ of the time-partial derivative $\partial_t V(t, x)$ at $t = 0$ of the velocity field $V(t)$.

As for first-order Eulerian semiderivatives, this first theorem reduces the study of second-order Eulerian semiderivatives to the autonomous case. So in §2.3.2 we will specialize to autonomous fields V and W in $\mathcal{D}^k(\mathbb{R}^N, \mathbb{R}^N)$ and give the equivalent of Hadamard's structure theorem for the canonical term.

2.3.1. Nonautonomous case. The basic framework introduced in §§2.1 and 2.2 has reduced the computation of the Eulerian semiderivative of $J(\Omega)$ to the computation of the derivative

$$j'(0) = dJ(\Omega; V(0)) \tag{21}$$

of the function

$$j(t) = J(\Omega_t(V)). \tag{22}$$

For $t \geq 0$, we naturally obtain

$$j'(t) = dJ(\Omega_t(V); V(t)). \tag{23}$$

This suggests the following definition.

DEFINITION 2.3. Let V and W belong to $\mathcal{L}$ and assume that for all $t \in [0, \tau]$, $dJ(\Omega_t(W); V(t))$ exists at $\Omega_t(W) = T_t(W)(\Omega)$ in the direction $V(t)$. The functional J is said to have a *second-order Eulerian semiderivative* at Ω in the directions (V, W) if the following limit exists:

$$\lim_{t \searrow 0} [dJ(\Omega_t(W); V(t)) - dJ(\Omega; V(0))]/t. \tag{24}$$

Whenever it exists, it is denoted by $d^2J(\Omega; V; W)$.

In this definition, the domain Ω is transformed into the domain $\Omega_t(W)$ under the action of the velocity field W. The vector function V appears in the expressions of $dJ(\Omega_t(W); V(t))$ and $dJ(\Omega; V(0))$, but does not contribute to the deformation of Ω.

Remark 2.5. This last definition is compatible with the second-order expansion of $j(t)$ with respect to t around $t = 0$:

$$j(t) \cong j(0) + tj'(0) + \frac{t^2}{2} j''(0), \tag{25}$$

where

$$j''(0) = d^2J(\Omega; V; V). \tag{26}$$

Remark 2.6. It is easy to construct simple examples with autonomous fields V and W showing that $d^2J(\Omega; V; W) \neq d^2J(\Omega; W; V)$ (cf. Delfour and Zolésio [8]).

The next theorem is the analogue of Theorem 2.3 and provides the canonical structure of the second-order Eulerian semiderivative.

THEOREM 2.5. *Let Ω be a domain in $\mathbb{R}^N$ and $m \geq 0$ and $\ell \geq 0$ be integers. Assume that*

(i) $\forall V \in \vec{\mathcal{V}}^{m+1,\ell}$, $\forall W \in \vec{\mathcal{V}}^{m,\ell}$, $d^2J(\Omega; V; W)$ *exists*;

(ii) $\forall W \in \vec{\mathcal{V}}^{m,\ell}$, $\forall t \in [0, \tau]$, *J has a shape gradient at $\Omega_t(W)$ of order ℓ*;

(iii) $\forall U \in \mathcal{V}^{\ell}$, *the map*

$$W \mapsto d^2J(\Omega; U; W) : \vec{\mathcal{V}}^{m,\ell} \to \mathbb{R} \tag{27}$$

is continuous. Then for all V in $\vec{\mathcal{V}}^{m+1,\ell}$ and all Win $\vec{\mathcal{V}}^{m,\ell}$

$$d^2J(\Omega; V; W) = d^2J(\Omega; V(0); W(0)) + dJ(\Omega; \dot{V}(0)), \tag{28}$$

where

$$\dot{V}(0)(x) = \lim_{t \searrow 0} [V(t, x) - V(0, x)]/t. \tag{29}$$

2.3.2. Autonomous case.

DEFINITION 2.4. Let Ω be a domain in $\mathbb{R}^N$.

(i) The functional $J(\Omega)$ is said to be *shape differentiable* at Ω if

$$\forall V, \quad \forall W \text{ in } \mathcal{D}(\mathbb{R}^N, \mathbb{R}^N), \quad d^2J(\Omega; V; W) \tag{30}$$

exist and the map

$$(V, W) \mapsto d^2J(\Omega; V; W) : \mathcal{D}(\mathbb{R}^N, \mathbb{R}^N) \times \mathcal{D}(\mathbb{R}^N, \mathbb{R}^N) \to \mathbb{R} \tag{31}$$

is bilinear and continuous. We denote by h the bilinear and continuous map (31).

(ii) Denote by $H(\Omega)$ the continuous linear map on the tensor product $\mathcal{D}(\mathbb{R}^N, \mathbb{R}^N) \otimes \mathcal{D}(\mathbb{R}^N, \mathbb{R}^N)$, associated with h:

$$(32) \qquad d^2J(\Omega; V; W) = \langle H(\Omega), V \otimes W\rangle = h(V, W),$$

where $V \otimes W$ is the tensor product of V and W defined as

$$(33) \qquad (V \otimes W)_{ij}(x, y) = V_i(x)W_j(y), \quad 1 \leq i, \quad j \leq N,$$

and $V_i(x)$ (respectively, $W_j(y)$) is the ith (respectively jth) component of the vector V(respectively W) (cf. Schwartz's [2] kernel theorem and Gelfand and Vilenkin [1]). $H(\Omega)$ will be called the *shape Hessian* of J at Ω.

(iii) When there exists a finite integer $\ell \geq 0$ such that $H(\Omega)$ is continuous on $\mathcal{D}^\ell(\mathbb{R}^N, \mathbb{R}^N) \otimes \mathcal{D}^\ell(\mathbb{R}^N, \mathbb{R}^N)$, we say that $H(\Omega)$ is of finite order ℓ.

THEOREM 2.6. *Let Ω be an open domain in $\mathbb{R}^N$ with boundary Γ and assume that J is twice shape differentiable at Ω.*

(i) *$H(\Omega)$ (that is, h) has support in $\Gamma \times \Gamma$. Moreover, when the support of $H(\Omega)$ is compact the order of $H(\Omega)$ is finite.*

(ii) *If $H(\Omega)$ is of finite order ℓ, $\ell \geq 0$, and Ω is an open domain with boundary Γ in $C^{\ell+1}$, then there exists a continuous linear map $h(\Omega)$ on the tensor product $\mathcal{D}^\ell(\Gamma, \mathbb{R}^N) \otimes \mathcal{D}^\ell(\Gamma)$ such that*

$$(34) \qquad d^2J(\Omega); V; W) = \langle h(\Omega), (\gamma_\Gamma V) \otimes ((\gamma_\Gamma W) \bullet n)\rangle,$$

where $\gamma_\Gamma V$ is the trace of V on Γ; $(\gamma_\Gamma V) \otimes ((\gamma_\Gamma W) \bullet n)$ is defined as the tensor product

$$(35) \quad ((\gamma_\Gamma V) \otimes (\gamma_\Gamma W) \bullet n))_i(x, y) = (\gamma_\Gamma V_i)(x)((\gamma_\Gamma W) \bullet n)(y), \quad x, y \in \Gamma, \quad 1 \leq i \leq N;$$

$V_i(x)$ is the ith component of $V(x)$; and

$$(36) \qquad ((\gamma_\Gamma W) \bullet n)(y) = (\gamma_\Gamma W)(y) \bullet n(y), \qquad y \in \Gamma.$$

Remark 2.7. Finally, under the assumptions of Theorems 2.5 and 2.6,

$$(37) \quad d^2J(\Omega; V; W) = \langle h(\Omega), (\gamma_\Gamma V(0)) \otimes ((\gamma_\Gamma W(0)) \bullet n)\rangle + \langle g(\Omega), (\gamma_\Gamma \dot{V}(0)) \bullet n\rangle$$

for all V in $\vec{\mathcal{V}}^{m+1,\ell}$ and W in $\vec{\mathcal{V}}^{m,\ell}$.

2.4. Comparison with methods of perturbation of the identity. At this juncture it is instructive to compare first- and second-order Eulerian semiderivatives obtained by the velocity (speed) method with those obtained by first- and second-order perturbations of the identity, that is, when the transformations T_t are specified a priori by

$$(38) \qquad T_t(X) = X + tU(X) + \frac{t^2}{2}A(X), \qquad X \in \mathbb{R}^N,$$

where U and A are transformations of $\mathbb{R}^N$ verifying the hypotheses of Theorem 2.2. The transformation T_t in (38) is a *second-order* perturbation when $A \neq 0$ and a *first-order* perturbation when $A = 0$.

According to Theorem 2.2, first- and second-order Eulerian semiderivatives associated with (38) can be equivalently obtained by applying the velocity (speed) method to the nonautonomous velocity fields V_{UA} given by (9). So when J verifies the hypotheses of Theorem 2.4,

$$(39) \qquad dJ(\Omega; V_{UA}) = dJ(\Omega; V_{UA}(0)) = dJ(\Omega; U),$$

where we have used the first part of Remark 2.2, which says that

$$V_{UA}(0) = U \text{ and } \overset{\bullet}{V}_{UA}(0) = A - [DU]U. \tag{40}$$

Similarly, if V_{WB} is another velocity field corresponding to

$$T_t(X) = X + tW(X) + \frac{t^2}{2}B(X),\ X \in \mathbb{R}^N, \tag{41}$$

where W and B verify the condition of Theorem 2.2, then when J verifies the assumptions of Theorem 2.6,

$$d^2J(\Omega; V_{UA}; V_{WB}) = d^2J(\Omega; V_{UA}(0); V_{WB}(0)) + dJ(\Omega; \overset{\bullet}{V}_{UA}(0)) \tag{42}$$

and

$$d^2J(\Omega; V_{UA}; V_{WB}) = d^2J(\Omega; U; W) + dJ(\Omega; A - [DU]U). \tag{43}$$

Expressions (39) and (43) are to be compared with the following expressions obtained by the velocity (speed) method for the two autonomous vector fields U and W:

$$dJ(\Omega; U) \quad \text{and} \quad d^2J(\Omega; U; W). \tag{44}$$

For the shape gradient the two expressions coincide; for the shape hessian we recognize the bilinear term in (43) and (44) but the two expressions differ by the term

$$dJ(\Omega; A - [DU]U). \tag{45}$$

Even for a first-order perturbation ($A = 0$), we have a quadratic term in U.

This situation is analogous to the classical problem of defining second-order derivatives on a manifold. The term (45) would correspond to the connection while the bilinear term $d^2J(\Omega; V; W)$ would be the candidate for the *canonical* second-order shape derivative. In this context, we will refer to the corresponding distribution $H(\Omega)$ as the *canonical shape Hessian*. All other second-order shape derivatives are obtained from $H(\Omega)$ by adding the gradient term $g(\Omega)$ acting as the appropriate acceleration field (connection).

Remark 2.8. The method of perturbation of the identity can be made *more canonical* by using the following family of transformations:

$$T_t(X) = X + tU(X) + \frac{t^2}{2}(A + [DU]U), \tag{46}$$

which yields

$$dJ(\Omega; U) \tag{47}$$

for the gradient and

$$d^2J(\Omega; U; W) + dJ(\Omega; A) \tag{48}$$

for the Hessian, where for a first-order perturbation ($A = 0$), the second term disappears.

Remark 2.9. Denote by $\mathcal{A}^{\ell+1}$ the set of all open domains Ω in $\mathbb{R}^N$ with a boundary Γ that is $C^{\ell+1}$, $\ell \geq 0$. Assume that there exists a domain Ω^* in $\mathcal{A}^{\ell+1}$ that minimizes a given domain functional $J(\Omega)$ over all Ω in $\mathcal{A}^{\ell+1}$. Then if J is twice shape differentiable for all Ω in $\mathcal{A}^{\ell+1}$, Ω^* verifies the following necessary conditions:

$$dJ(\Omega^*; V) = 0 \quad \forall V \in \mathcal{D}(\mathbb{R}^N, \mathbb{R}^N), \tag{49}$$

$$d^2J(\Omega^*; W; W) \geq 0 \quad \forall W \in \mathcal{D}(\mathbb{R}^N, \mathbb{R}^N). \tag{50}$$

Obviously, necessary conditions are not, in general, sufficient conditions. This fact is well known in optimization problems over vector spaces. For shapes the situation is even more delicate since a space of domains is not a vector space, and traditional concepts such as convexity are more difficult to formalize and use. We also know that, always in general, optimal domains are not necessarily smooth. Microstructures or domains with fractal boundaries naturally occur as in the optimization of the thickness of a plate (cf. Cheng and Olhoff [1]). This type of phenomenon is similar to the one encountered in chattering control.

3. A saddle point formulation of the Dirichlet problem. Let Ω be a bounded open domain in $\mathbb{R}^N$ with a sufficiently smooth boundary Γ. Let f and g be two fixed functions in $H^{1/2+\epsilon}(\mathbb{R}^N)$ and $H^{2+\epsilon}(\mathbb{R}^N)$, respectively, for some arbitrary small $\epsilon > 0$. Consider the solution y in $H^2(\Omega)$ to the nonhomogeneous Dirichlet boundary value problem:

$$-\Delta y = f \quad \text{in } \Omega, \qquad y = g \quad \text{on } \Gamma. \tag{1}$$

Our objective is to transform this problem into finding the saddle point of a volume Lagrangian functional. This technique can be applied to other boundary value problems with Dirichlet conditions.

Note that y is also the solution of the weak equation

$$\int_\Omega (\Delta y + f)\psi \, dx + \int_\Gamma (y - g)\mu \, d\Gamma = 0 \tag{2}$$

for all ψ in $H^2(\Omega)$ and μ in $H^{1/2}(\Gamma)$, since the corresponding continuous convex-concave functional

$$L(\phi,\ \psi,\ \mu) = \int_\Omega (\Delta\phi + f)\psi \, dx + \int_\Gamma (\phi - g)\mu \, d\Gamma \tag{3}$$

has a unique saddle point $(\hat{\phi}, \hat{\psi}, \hat{\mu})$, which is completely characterized by the equations

$$\Delta\hat{\phi} + f = 0 \quad \text{in } \Omega, \tag{4}$$

$$\hat{\phi} - g = 0 \quad \text{in } \Gamma, \tag{5}$$

$$\int_\Omega \Delta\phi \ \hat{\psi} \, dx + \int_\Gamma \phi\hat{\mu} \, d\Gamma = 0 \quad \forall\phi \in H^2(\Omega), \tag{6}$$

where the last equation yields

$$\Delta\hat{\psi} = 0 \quad \text{in } \Omega, \qquad \hat{\psi} = 0 \quad \text{on } \Gamma, \quad \text{and} \quad \hat{\mu} = \frac{\partial\hat{\psi}}{\partial n} \quad \text{on } \Gamma. \tag{7}$$

The proof of this can be found in Ekeland and Temam [1, Prop. 1.6]. Of course, this implies that the saddle point is unique and given by

$$(\hat{\phi}, \hat{\psi}, \hat{\mu}) = (y, 0, 0). \tag{8}$$

The purpose of the above computation was to characterize the multiplier $\hat{\mu}$

$$\hat{\mu} = \frac{\partial\hat{\psi}}{\partial n} \quad \text{on } \Gamma, \tag{9}$$

to rewrite the previous functional as a function of two variables instead of three:

$$L(\phi,\psi)=\int_\Omega(\Delta\phi+f)\psi\,dx+\int_\Gamma(\phi-g)\frac{\partial\psi}{\partial n}\,d\Gamma, \tag{10}$$

for (ϕ,ψ) in $H^2(\Omega)\times H^2(\Omega)$. It is also advantageous for shape problems to get rid of boundary integrals whenever it is possible. So noting that

$$\int_\Gamma(\phi-g)\frac{\partial\psi}{\partial n}\,d\Gamma=\int_\Omega \operatorname{div}[(\phi-g)\nabla\psi]\,dx, \tag{11}$$

we finally use the functional

$$L(\phi,\psi)=\int_\Omega\{(\Delta\phi+f)\psi+(\phi-g)\Delta\psi+\nabla(\phi-g)\bullet\nabla\psi\}\,dx \tag{12}$$

on $H^2(\Omega)\times H^2(\Omega)$. It is readily seen that it has a unique saddle point $(\hat\phi,\hat\psi)$ in $H^2(\Omega)\times H^2(\Omega)$, which is completely characterized by the following saddle point equations:

$$\Delta\hat\phi+f=0\quad\text{in }\Omega,\qquad \hat\psi=g\quad\text{on }\Gamma,\quad \Delta\hat\psi=0\quad\text{in }\Omega,\quad \hat\psi=0\quad\text{on }\Gamma. \tag{13}$$

4. Shape gradient for the Dirichlet problem.

4.1. Formulation and formal computations. Consider the cost function

$$J(\Omega)=\frac{1}{2}\int_\Omega|y(\Omega)-y_d|^2\,dx \tag{1}$$

associated with the solution $y=y(\Omega)$ of the Dirichlet problem (3.1) and the given function y_d in $H^{1/2+\epsilon}(\mathbb{R}^N)$ for some arbitrary fixed $\epsilon>0$.

As in §3, we reformulate this problem as the saddle point of a functional by introducing the Lagrangian

$$G(\Omega,\phi,\psi)=\frac{1}{2}\int_\Omega|\phi-y_d|^2\,dx+\int_\Omega\{(\Delta\phi+f)\psi+(\phi-g)\Delta\psi+\nabla(\phi-g)\bullet\nabla\psi\}\,dx \tag{2}$$

on $H^2(\Omega)\times H^2(\Omega)$. It is readily seen that $G(\Omega,\cdot,\cdot)$ has a unique saddle point $(\hat\phi,\hat\psi)$, which is completely characterized by the following saddle point equations:

$$\Delta\hat\phi+f=0\quad\text{in }\Omega,\quad \hat\phi=g\quad\text{on }\Gamma, \tag{3}$$

$$\int_\Omega\{(\hat\phi-y_d)\phi+\Delta\phi\hat\psi+\phi\Delta\hat\psi+\nabla\phi\bullet\nabla\hat\psi\}\,dx=0\quad\forall\ \phi\in H^2(\Omega). \tag{4}$$

The last equation is equivalent to

$$\int_\Omega[(\hat\phi-y_d)+\Delta\hat\psi]\phi\,dx+\int_\Gamma\frac{\partial\phi}{\partial n}\hat\psi\,d\Gamma=0\quad\forall\phi\in H^2(\Omega) \tag{5}$$

or

$$\Delta\hat\psi+(\hat\phi-y_d)=0\text{ in }\Omega,\quad\hat\psi=0\text{ on }\Gamma, \tag{6}$$

by using the theorem on the surjectivity of the trace. In the folllowing, we will use the notation (y,p) for the saddle point $(\hat\phi,\hat\psi)$. As a result, we have

$$J(\Omega)=\operatorname*{Min}_{\phi\in H^2(\Omega)}\ \operatorname*{Max}_{\psi\in H^2(\Omega)}\ G(\Omega,\phi,\psi). \tag{7}$$

We will now use the above Lagrangian formulation combined with the velocity method (cf. Céa [1]–[3]; Zolésio [1], [2], Delfour and Zolésio [1]–[4], [7]–[9]) to compute the shape gradient of $J(\Omega)$. Recall that the domain Ω is perturbed by a velocity vector field V that induces a family of homeomorphisms (cf. §2.1)

$$T_t : \mathbb{R}^N \to \mathbb{R}^N, \qquad T_t(X) = x(t), \tag{8}$$

which transforms the domain Ω into the new domains

$$\Omega_t = T_t(\Omega), \quad t \in [0,\tau]. \tag{9}$$

The Shape semiderivative is defined as (cf. §2.2)

$$dJ(\Omega; V) = \lim_{t\searrow 0}[J(\Omega_t) - J(\Omega)]/t \tag{10}$$

whenever the limit exists. For t in $[0,\tau]$

$$J(\Omega_t) = \operatorname*{Min}_{\phi\in H^2(\Omega_t)} \ \operatorname*{Max}_{\psi\in H^2(\Omega_t)} \ G(\Omega_t, \phi, \psi) \tag{11}$$

and the technical difficulty arises from the time dependence of the underlying function spaces. There are two methods to get around the time dependence in the underlying function spaces (cf. Delfour and Zolésio [1], [2]):

– the *function space parametrization* and

– the *function space embedding.*

In the first case, we parametrize the functions in $H^2(\Omega_t)$ by elements of $H^2(\Omega)$ through the transformation

$$\phi \mapsto \phi \circ T_t^{-1} \ : \ H^2(\Omega) \to H^2(\Omega_t), \tag{12}$$

where "$\circ$" denotes the composition of the two maps and we introduce the *parametrized Lagrangian*,

$$\tilde{G}(t,\phi,\psi) = G(T_t(\Omega), \phi \circ T_t^{-1}, \psi \circ T_t^{-1}) \tag{13}$$

on $H^2(\Omega) \times H^2(\Omega)$. In the function space embedding method, we introduce a large enough fixed domain, D which contains all the transformations $\{\Omega_t : 0 \le t \le \tau\}$ of Ω.

In this paper, we will use the function space embedding method with $D = \mathbb{R}^N$

$$J(\Omega_t) = \operatorname*{Min}_{\Phi\in H^2(\mathbb{R}^N)} \ \operatorname*{Max}_{\Psi\in H^2(\mathbb{R}^N)} \ G(\Omega_t, \Phi, \Psi). \tag{14}$$

Capital letters will be used for the functions on $\mathbb{R}^N$ and lowercase letters for the functions on Ω or Ω_t. As could be expected, the price to pay for the use of this method is that the set of saddle points

$$S(t) = X(t) \times Y(t) \ \subset H^2(\mathbb{R}^N) \times H^2(\mathbb{R}^N) \tag{15}$$

is not a singleton anymore since

$$X(t) = \{\Phi \in H^2(\mathbb{R}^N) : \Phi|_{\Omega_t} = y_t\}, \tag{16}$$

$$Y(t) = \{\Psi \in H^2(\mathbb{R}^N) : \Psi|_{\Omega_t} = p_t\}, \tag{17}$$

where (y_t, p_t) is the unique solution in $H^2(\Omega_t) \times H^2(\Omega_t)$ to the previous saddle point equations on Ω_t

$$\Delta y_t + f = 0 \quad \text{in} \quad \Omega_t, \qquad y_t = g \quad \text{on } \Gamma_t, \tag{18}$$

$$\Delta p_t + (y_t - y_d) = 0 \ \text{ in } \ \Omega_t, \qquad p_t = 0 \quad \text{on } \ \Gamma_t. \tag{19}$$

We are now ready to apply the theorem of Correa and Seeger [1], which says that under appropriate hypotheses (to be checked in the next section)

$$dJ(\Omega; V) = \operatorname*{Min}_{\Phi \in X(0)} \operatorname*{Max}_{\Psi \in Y(0)} \partial_t G(\Omega_t, \Phi, \Psi). \tag{20}$$

Since we have already characterized $X(0)$ and $Y(0)$, we only need to compute the partial derivative of

(21)

$$G(\Omega_t, \Phi, \Psi) = \int_{\Omega_t} \left\{ \frac{1}{2} |\Phi - y_d|^2 + (\Delta\Phi + f)\Psi + (\Phi - g)\Delta\Psi + \nabla(\Phi - g) \bullet \nabla\Psi \right\} dx.$$

If we assume that Ω_t is sufficiently smooth, then

$$f, y_d \in H^{1/2+\epsilon}(\mathbb{R}^N), \quad g \in H^{2+\epsilon}(\mathbb{R}^N) \Rightarrow y, \quad p \in H^{\frac{5/2+\epsilon}{(}}\Omega) \tag{22}$$

and we can choose to consider our saddle points $S(t)$ in $H^{5/2+\epsilon}(\mathbb{R}^N) \times H^{5/2+\epsilon}(\mathbb{R}^N)$ rather than $H^2(\mathbb{R}^N) \times H^2(\mathbb{R}^N)$. If Φ and Ψ belong to $H^{5/2+\epsilon}(\mathbb{R}^N)$, then

$$\begin{aligned} &\partial_t G(\Omega_t, \Phi, \Psi) \\ &= \int_{\Gamma_t} \{\frac{1}{2}(\Phi - y_d)^2 + (\Delta\Phi + f)\Psi + (\Phi - g)\Delta\Psi + \nabla(\Phi - g) \bullet \nabla\Psi\} V \bullet n_t \, d\Gamma_t. \end{aligned} \tag{23}$$

This expression is an integral over the boundary Γ, which will not depend on Φ and Ψ outside of $\bar{\Omega}$. As a result the Min and the Max can be dropped in expression (20), which reduces to

(24)

$$dJ(\Omega; V) = \int_{\Gamma} \left\{ \frac{1}{2}(y - y_d)^2 + (\Delta y + f)p + (y - g)\Delta p + \nabla(y - g) \bullet \nabla p \right\} V \bullet n \, d\Gamma.$$

But

$$p = 0 \quad \text{and} \quad y - g = 0 \quad \Rightarrow \quad \nabla p = \frac{\partial p}{\partial n} n, \quad \nabla(y - g) = \frac{\partial}{\partial n}(y - g)\, n \quad \text{on } \Gamma \tag{25}$$

and finally

$$dJ(\Omega; V) = \int_{\Gamma} \left\{ \frac{1}{2}(g - y_d)^2 + \frac{\partial}{\partial n}(y - g)\frac{\partial p}{\partial n} \right\} V \bullet n \, d\Gamma. \tag{26}$$

4.2. Verification of the hypotheses. As we have seen, the computation of the shape gradient is both quick and easy. We now turn to the step-by-step verification of the hypotheses of the underlying theorem. Many of the constructions given below are "canonical" and can be repeated for different problems in different contexts.

THEOREM 4.1 (Correa and Seeger [1]). *Let $\tau > 0$, the sets X and Y, and the functional $L : [0, \tau] \times X \times Y \to \mathbb{R}$ be given. Denote by*

$$S(t) = X(t) \times Y(t) \subset X \times Y \tag{27}$$

the set of saddle points of the functional $L(t, \cdot\, , \cdot)$ on $X \times Y$. Assume that

$$\forall t \in [0, \tau], \quad S(t) \neq \varnothing \tag{H1}$$

and that

(H2) $$\forall(x,y) \in \left[X(0) \times \bigcup_{0\le t\le\tau} Y(t)\right] \cup \left[\bigcup_{0\le t\le\tau} X(t) \times Y(0)\right],$$
$$\partial_t L(t,x,y) \textit{ exists on } [0,\tau].$$

Moreover, assume that there exist topologies $\mathcal{T}_X$ on X and $\mathcal{T}_Y$ on Y such that for all sequences $t_n \to 0$ as $n \to \infty$, $0 \le t_n \le \tau$, there exist $(x_0, y_0) \in S(0)$ and a subsequence of $\{t_n\}$, still denoted $\{t_n\}$, such that

(H3) $\forall n, \quad \exists (x_n, y_n) \in S(t_n) \quad$ *and* $\quad (x_n, y_n) \to (x_0, y_0) \quad$ *in* $\mathcal{T}_X \times \mathcal{T}_Y \quad$ *as* $n \to \infty$,

(H4) $$\forall y \quad \textit{in } Y(0)_2 \quad \liminf_{\substack{t\searrow 0\\ n\to\infty}} \partial_t L(t, x_n, y) \ge \partial_t L(0, x_0, y)$$
$$(\textit{resp.}\, \forall x \in X(0),\ \limsup_{\substack{t\searrow 0\\ n\to\infty}} \partial_t L(t, x, y_n) \le \partial_t L(0, x, y_0)).$$

Then the function
$$g(t) = \underset{x\in X}{\text{Min}}\ \underset{y\in Y}{\text{Max}}\ L(t,x,y)$$
on $[0,\tau]$ has a semiderivative at $t = 0$ given by
$$dg(0) = \lim_{t\searrow 0}[g(t) - g(0)]/t$$
$$= \underset{x\in X(0)}{\text{Inf}}\ \underset{y\in Y(0)}{\text{Sup}}\ \partial_t L(0,x,y) = \underset{y\in Y(0)}{\text{Sup}}\ \underset{x\in X(0)}{\text{Inf}}\ \partial_t L(0,x,y).$$

Let y_d and $f \in H^1(\mathbb{R}^N)$ and $g \in H^{5/2}(\mathbb{R}^N)$ so that

(28) $$X = Y = H^3(\mathbb{R}^N).$$

The saddle points $S(t) = X(t) \times Y(t)$ are given by

(29) $$X(t) = \{\Phi \in X : \Phi|_{\Omega_t} = y_t\},$$

(30) $$Y(t) = \{\Psi \in Y : \Psi|_{\Omega_t} = p_t\}.$$

The sets $X(t)$ and $Y(t)$ are not empty since it is always possible to construct a continuous linear extension

(31) $$\Pi^m : H^m(\Omega) \to H^m(\mathbb{R}^N)$$

for each $m \ge 1$. For instance, with $m = 1$ and a boundary Γ, which is $W^{1,\infty}$, see Agmon, Douglis, and Nirenberg [1], [2] and, for $m > 1$, see Babić [1] (cf. also Nečas [1]). Using this Π^m, we then define the following extension:

(32) $$\Pi_t^m : H^m(\Omega_t) \to H^m(\mathbb{R}^N),$$

(33) $$\Pi_t^m(\phi) = [\Pi^m(\phi \circ T_t)] \circ T_t^{-1}.$$

In the following, m is fixed and equal to 3, so we will drop the superscript m and define the extensions

(34) $$Y_t = \Pi_t y_t, \qquad P_t = \Pi_t p_t$$

of y_t and p_t, respectively. Hence,

$$Y_t \in X(t) \quad \text{and} \quad P_t \in Y(t) \Rightarrow S(t) \neq \emptyset. \tag{35}$$

So condition (H1) is verified. Condition (H2) follows from the hypotheses on f, y_d, and g. To check conditions (H3) and (H4), we need two general theorems that can be used in various contexts and problems.

THEOREM 4.2. *For $V \in \mathcal{D}^1(\mathbb{R}^N, \mathbb{R}^N)$ and $\Phi \in L^2(\mathbb{R}^N)$,*

$$\lim_{t \searrow 0} \Phi \circ T_t = \Phi \quad \textit{and} \quad \lim_{t \searrow 0} \Phi \circ T_t^{-1} = \Phi \quad \textit{in} L^2(\mathbb{R}^N). \tag{36}$$

Proof. (i) The space $\mathcal{D}(\mathbb{R}^N)$ of infinitely continuously differentiable functions with compact support in $\mathbb{R}^N$ is dense in $L^2(\mathbb{R}^N)$. So given $\epsilon > 0$, there exists Φ_ϵ in $\mathcal{D}(\mathbb{R}^N)$such that

$$||\Phi - \Phi_\epsilon||_{L^2}^2 < \epsilon^2 / \max \{J_t^{-1} : 0 \leq t \leq \tau\}.$$

Hence,

$$||\Phi \circ T_t - \Phi|| \leq ||\Phi_\epsilon \circ T_t - \Phi_\epsilon|| + ||\Phi \circ T_t - \Phi_\epsilon \circ T_t|| + ||\Phi - \Phi_\epsilon||. \tag{37}$$

But

$$\forall t \in [0, \tau], \quad \int_{\mathbb{R}^N} |\Phi \circ T_t - \Phi_\epsilon \circ T_t|^2 \, dx = \int_{\mathbb{R}^N} |\Phi - \Phi_\epsilon|^2 J_t^{-1} \, dx \leq \epsilon^2.$$

So the last two terms in (37) are less than 2ϵ. It remains to evaluate the first term for a fixed function Φ_ϵ with compact support K in $\mathbb{R}^N$. Recall that, since $\Phi_\epsilon = 0$ on the boundary ∂K of K, $T_t(K) = K$ for all t in $[0, \tau]$ (use Nagumo's [1] theorem twice as in the proof of Theorem 2.4(i)). Moreover, by compactness of K, Φ_ϵ is uniformly continuous on $\mathbb{R}^N$ and

$$\exists \delta > 0, \quad \forall x, y \in \mathbb{R}^N, \quad |x - y| < \delta \Rightarrow |\Phi_\epsilon(y) - \Phi_\epsilon(x)| < \epsilon / m(K)^{1/2}.$$

T_t is also uniformly continuous on K and

$$\exists \eta > 0, \quad \forall t, \ 0 \leq t < \eta, \quad \forall x \in K, \quad |T_t x - x| < \delta.$$

By construction

$$\text{supp } (\Phi_\epsilon \circ T_t) = T_t \text{ (supp } \Phi_\epsilon) \subset K$$

and

$$\Phi_\epsilon = 0 \text{ and } \Phi_\epsilon \circ T_t = 0 \text{ outside of } K.$$

Finally,

$$\int_{\mathbb{R}^N} |\Phi_\epsilon(T_t x) - \Phi_\epsilon(x)|^2 \, dx = \int_K |\Phi_\epsilon(T_t x) - \Phi_\epsilon(x)|^2 \, dx \leq \epsilon^2$$

and this implies that

$$\forall \epsilon > 0, \quad \exists \eta > 0, \quad \forall t, \ 0 \leq t \leq \eta, \quad ||\Phi \circ T_t - \Phi||_{L^2(\mathbb{R}^N)} \leq 3\epsilon.$$

(ii) For the second part of (36) we make a change of variable and use the result of part (i)

$$\int_{\mathbb{R}^N} |\Phi \circ T_t^{-1} - \Phi|^2 \, dx = \int_{\mathbb{R}^N} |\Phi - \Phi \circ T_t|^2 J_t \, dx \leq \epsilon^2.$$

This completes the proof of Theorem 4.2. □

COROLLARY. *Under the assumptions of Theorem* 4.2 *for* $m \geq 1$, V *in* $\mathcal{D}^m(\mathbb{R}^N, \mathbb{R}^N)$, *and* $\Phi \in H^m(\mathbb{R}^N)$,

$$(38) \qquad \lim_{t \searrow 0} \Phi \circ T_t = \Phi \quad \textit{and} \quad \lim_{t \searrow 0} \Phi \circ T_t^{-1} = \Phi \quad \textit{in} \quad H^m(\mathbb{R}^N).$$

Remark 4.1. In fact, for $m \geq 1$ and $V \in \mathcal{D}^m(\mathbb{R}^N, \mathbb{R}^N)$ the transformation

$$(39) \qquad S(t)\Phi = \Phi \circ T_t, \quad \forall \Phi \in H^m(\mathbb{R}^N), \quad \forall t, \quad 0 \leq t \leq \tau,$$

defines a strongly continuous semigroup of class C_0 on $H^m(\mathbb{R}^N)$ with infinitesimal generator

$$\mathcal{A}\Phi = \nabla\Phi \bullet V, \qquad \mathcal{D}(\mathcal{A}) = \{\Phi \in H^m(\mathbb{R}^N) : \nabla\Phi \bullet V \in H^m(\mathbb{R}^N)\}.$$

THEOREM 4.3. *Under the assumptions of Theorem* 4.2,

$$(40) \qquad y^t \longrightarrow y^0 \quad \textit{in} \quad H^m(\Omega) - \textit{strong (respectively, weak)}$$

implies that

$$Y_t \longrightarrow Y_0 \quad \textit{in} \quad H^m(\mathbb{R}^N) - \textit{strong (respectively, weak)}.$$

Proof. The strong case is obvious. We prove the weak case for $m = 0$. By definition,

$$Y_t = (\Pi y^t) \circ T_t^{-1}$$

and for all Φ in $L^2(\mathbb{R}^N)$, we consider

$$\int_{\mathbb{R}^N} Y_t \ \Phi \, dx = \int_{\mathbb{R}^N} (\Pi y^t) \circ T_t^{-1} \ \Phi \, dx = \int_{\mathbb{R}^N} \Pi y^t \ \Phi \circ T_t \ J_t \, dx.$$

We have shown in Theorem 4.2 that

$$\Phi \circ T_t \longrightarrow \Phi \quad \text{in} \quad L^2(\mathbb{R}^N)\text{-} - \textit{strong} \ .$$

In addition, $J_t \longrightarrow 1$ and by linearity and continuity of Π

$$\Pi y^t \longrightarrow \Pi y \quad \text{in} \quad L^2(\mathbb{R}^N) - \text{weak}.$$

Hence,

$$\forall \Phi \in L^2(\mathbb{R}^N), \qquad \int_{\mathbb{R}^N} Y_t \ \Phi \, dx \longrightarrow \int_{\mathbb{R}^N} \Pi y \ \Phi \, dx = \int_{\mathbb{R}^N} Y_0 \ \Phi \, dx.$$

This proves the weak convergence. □

To verify condition (H3), we transform (y_t, p_t) on Ω_t to $(y^t, p^t) = (y_t \circ T_t, p_t \circ T_t)$ on Ω. The pair (y^t, p^t) is the transported pair of solutions from Ω_t to Ω. It is the unique solution in $H^1(\Omega) \times H^1(\Omega)$ of the system

$$(41) \qquad -\text{div}[A(t)\nabla y^t] = J_t \ f \circ T_t \quad \text{in} \quad \Omega, \qquad y^t = g \circ T_t \quad \text{on} \quad \Gamma,$$

$$(42) \qquad -\text{div}[A(t)\nabla p^t] = J_t(y^t - y_d \circ T_t) \ \text{ in } \ \Omega, \qquad p^t = 0 \ \text{ on } \ \Gamma,$$

where

$$(43) \qquad A(t) = J_t(DT_t)^{-1*}(DT_t)^{-1}, \qquad J_t = |\text{det} DT_t|,$$

DT_t is the Jacobian matrix of T_t and $(DT_t)^{-1*}$ is the transposed form of $(DT_t)^{-1}$.

For sufficiently smooth domains Ω and vector fields V, the pair $\{y^t, p^t\}$ is bounded in $H^1(\Omega) \times H^1(\Omega)$ as t goes to zero . Since $H^1(\Omega)$ is a Hilbert space, we can extract weakly convergent subsequences to some $(\bar{y}, \bar{p})$ in $H^1(\Omega) \times H^1(\Omega)$. However, by linearity of the equation with respect to (y^t, p^t) and continuity of the coefficients with respect to t, the limit point $(\bar{y}, \bar{p})$ will coincide with (y^0, p^0), since the system has a unique solution at $t = 0$. Then we go back to the equation for y^t and y and show that the convergence is strong in $H^1(\Omega)$. Finally, by using the regularity of the data and the classical regularity theorems we show that $(y^t, p^t) \to (y, p)$ in $H^3(\Omega) \times H^3(\Omega)$.

For the verification of condition (H4), we go back to expression (3.23), which can be rewritten as a volume integral

$$\begin{aligned} &\partial_t G(\Omega_t, \Phi, \Psi) \\ (44) \quad &= \int_{\Omega_t} \operatorname{div}\left\{ \left[\frac{1}{2}(\Phi - y_d)^2 + (\Delta\Phi + f)\Psi + (\Phi - g)\Delta\Psi + \nabla(\Phi - g) \bullet \nabla\Psi \right] V \right\} dx \end{aligned}$$

for $(\Phi, \Psi) \in H^3(\mathbb{R}^N) \times H^3(\mathbb{R}^N)$. Now introduce the map

$$\begin{aligned} (\Phi, \Psi) \mapsto F(\Phi, \Psi) &= \left[\frac{1}{2}(\Phi - y_d)^2 + (\Delta\Phi + f)\Psi + (\Phi - g)\Delta\Psi + \nabla(\Phi - g) \bullet \nabla\Psi \right] V \\ &: H^3(\mathbb{R}^N) \times H^3(\mathbb{R}^N) \to (H^1(\mathbb{R}^N))^N. \end{aligned}$$

It is bilinear and continuous. Finally, the map

$$\begin{aligned} (45) \quad & (t, F) \mapsto \int_{\Gamma_t} F \circ n_t \, d\Gamma = \int_{\Omega_t} F \, dx = \int_{\Omega} (\operatorname{div} F) \circ T_t \; J_t^{-1} \, dx \\ & : [0, \tau] \times H^1(\mathbb{R}^N) \to \mathbb{R} \end{aligned}$$

is continuous. Then

$$(46) \qquad (t, \Phi, \Psi) \mapsto \partial_t G(\Omega_t, \Phi, \Psi) = \int_{\Gamma_t} F(\Phi, \Psi) \bullet n_t \, d\Gamma_t$$

is continuous and condition (H4) is verified. This completes the verification of the four conditions of Theorem 4.1. □

5. Shape Hessian for the Dirichlet problem.

5.1. Formulation and formal computations. We proceed as in §§3 and 4 and provide the mathematical justification in §5.2. For the second derivative, we need two autonomous vector fields V and W on $\mathbb{R}^N$ and the expression of the first derivative $dJ(\Omega_t(W); V)$, where $\Omega_t(W)$ is the perturbation of the domain Ω by the vector field W:

$$(1) \qquad dJ(\Omega_t(W); V) = \int_{\Omega_t(W)} \operatorname{div}\left\{ \left[\frac{1}{2}(g - y_d)^2 + \nabla(y_t - g) \bullet \nabla p_t \right] V \right\} dx,$$

where (y_t, p_t) are the unique solutions in $H^3(\Omega_t(W)) \times H^3(\Omega_t(W))$ to the equations

$$(2) \qquad \Delta y_t + f = 0 \quad \text{in } \Omega_t(W), \qquad y_t = g \quad \text{on } \Gamma_t(W),$$

$$(3) \qquad \Delta p_t + (y_t - y_d) = 0 \quad \text{in } \Omega_t(W), \qquad p_t = 0 \text{ on } \Gamma_t(W).$$

Then, we express (1) as a MinMax over a new Lagrangian:

$$(4) \qquad dJ(\Omega_t(W); V) = \underset{\Phi, \Psi \in H^3(\mathbb{R}^N)}{\operatorname{Min}} \; \underset{P, \Sigma \in H^2(\mathbb{R}^N)}{\operatorname{Max}} G(\Omega_t, \Phi, \Psi, P, \Sigma),$$

where $G = G(\Omega_t, \Phi, \Psi, P, \Sigma)$ is given by

$$G = \int_{\Omega_t} \left\{ \operatorname{div} \left[\frac{1}{2}(g - y_d)^2 + \nabla(\Phi - g) \bullet \nabla\Psi \right] V \right\} + [\Delta\Phi + f]P + (\Phi - g)\Delta P + \nabla(\Phi - g) \bullet \nabla P + [\Delta\Psi + \Phi - y_d]\Sigma + \Psi\Delta\Sigma + \nabla\Psi \bullet \nabla\Sigma \Big\} \, dx. \tag{5}$$

This new Lagrangian is affine in (P, Σ), but is not necessarily convex in (Φ, Ψ). However, it is semiconvex in (Φ, Ψ) and we will see that Correa and Seeger [1] will still apply to our special Lagrangian where the sets $X(t) \times Y(t)$,

$$X(t) \subset H^3(\mathbb{R}^N) \times H^3(\mathbb{R}^N), \tag{6}$$

$$Y(t) \subset H^2(\mathbb{R}^N) \times H^2(\mathbb{R}^N) \tag{7}$$

will be given by the usual "saddle point equations":

$$\int_{\Omega_t} [\Delta\hat{\Phi} + f]P + (\hat{\Phi} - g)\Delta P + \nabla(\hat{\Phi} - g) \bullet \nabla P \, dx = 0 \quad \forall P, \tag{8}$$

$$\int_{\Omega_t} [\Delta\hat{\Psi} + \hat{\Phi} - y_d]\Sigma + \hat{\Psi}\Delta\Sigma + \nabla\hat{\Psi} \bullet \nabla\Sigma \, dx = 0 \quad \forall\Sigma, \tag{9}$$

$$\int_{\Omega_t} \left[\operatorname{div}\{[\nabla(\hat{\Phi} - g) \bullet \nabla\Psi]V\} + \Delta\Psi\hat{\Sigma} + \Psi\Delta\hat{\Sigma} + \nabla\Psi \bullet \nabla\hat{\Sigma} \right] dx = 0 \quad \forall\Psi, \tag{10}$$

$$\int_{\Omega_t} \left[\operatorname{div}\{[\nabla\Phi \bullet \nabla\hat{\Psi}]V\} + \Delta\Phi\hat{P} + \Phi\Delta\hat{P} + \nabla\Phi \bullet \nabla\hat{P} + \Phi\hat{\Sigma} \right] dx = 0 \quad \forall\Phi. \tag{11}$$

It is obvious that (8) and (9) yield

$$\hat{\Phi}|_{\Omega_t} = y_t \quad \text{and} \quad \hat{\Psi}|_{\Omega_t} = p_t. \tag{12}$$

Similarly, (10) and (11) have solutions $(\hat{\Sigma}, \hat{P})$ in $H^2(\mathbb{R}^N) \times H^2(\mathbb{R}^N)$ such that

$$Y_t' = \hat{\Sigma}|_{\Omega_t}, \qquad P_t' = \hat{P}|_{\Omega_t} \tag{13}$$

are unique in $H^2(\Omega_t) \times H^2(\Omega_t)$ and solutions of

$$\Delta Y_t' = 0 \text{ in } \Omega_t(W), \qquad Y_t' = -\frac{\partial}{\partial n_t}(y_t - g)V \bullet n_t \quad \text{on } \Gamma_t(W), \tag{14}$$

$$\Delta P_t' = 0 \quad \text{in } \Omega_t(W), \qquad P_t' = -\frac{\partial p_t}{\partial n_t} V \bullet n_t \quad \text{on } \Gamma_t(W). \tag{15}$$

It can be shown that Y_t' and P_t' coincide with the "partial derivative" with respect to t of appropriate extensions of y_t and p_t from $\Omega_t(W)$ to $\mathbb{R}^N$.

Finally, the partial derivative of the Lagrangian G with respect to t is given by

$$\partial_t G = \int_{\Gamma_t} \left\{ \operatorname{div} \left\{ \left[\frac{1}{2}(g - y_d)^2 + \nabla(\Phi - g) \bullet \nabla\Psi \right] V \right\} + [\Delta\Phi + f]P + (\Phi - g)\Delta P + \nabla(\Phi - g) \bullet \nabla P + [\Delta\Psi + \Phi - y_d]\Sigma + \Psi\Delta\Sigma + \nabla\Psi \bullet \nabla\Sigma \right\} W \bullet n_t \, d\Gamma_t \tag{16}$$

for Φ, Ψ, P, Σ in $H^3(\mathbb{R}^N)$, y_d and f in $H^2(\mathbb{R}^N)$, and g in $H^{7/2}(\mathbb{R}^N)$. The immediate consequence of this computation is that y_t, p_t, Y_t', P_t' all belong to $H^3(\Omega_t)$. But Y_t', P_t' in $H^3(\Omega_t)$ require that y_t and p_t belong to $H^4(\Omega_t)$. This is precisely why we chose the above smoothness for y_d, f, and g.

Therefore, we must consider our saddle points $X(t) \times Y(t)$ in $(H^4(\mathbb{R}^N) \times H^4(\mathbb{R}^N)) \times (H^3(\mathbb{R}^N) \times H^3(\mathbb{R}^N))$,

$$X(t) = \{(\Phi, \Psi) \in H^4(\mathbb{R}^N) \times H^4(\mathbb{R}^N) : \Phi|_{\Omega_t} = y_t, \Psi|_{\Omega_t} = p_t\}, \tag{17}$$

$$Y(t) = \{(P, \Sigma) \in H^3(\mathbb{R}^N) \times H^3(\mathbb{R}^N) : P|_{\Omega_t} = P_t', \Sigma|_{\Omega_t} = Y_t'\}. \tag{18}$$

Finally, since $\partial_t G$ is a functional on Ω_t, it will only use the restriction to Ω_t of the various functions in $X(t) \times Y(t)$. Therefore, the Min and the Max can be removed and

$$\begin{aligned} d^2J(\Omega; V; W) = \int_\Gamma \Bigg\{ &\operatorname{div}\left\{\left[\frac{1}{2}(y - y_d)^2 + \frac{\partial}{\partial n}(y - g)\frac{\partial p}{\partial n}\right] V\right\} \\ &+ [\Delta y + f]P_V' + (y - g)\Delta P_V' + \nabla(y - g) \bullet \nabla P_V' \\ &+ [\Delta p + y - y_d]Y_V' + p\Delta Y_V' + \nabla p \bullet \nabla Y_V' \Bigg\} W \bullet n \, d\Gamma. \end{aligned} \tag{19}$$

But

$$\Delta y + f = 0, \quad y = g, \quad \Delta p + y - y_d = 0, \quad \text{and} \quad p = 0 \quad \text{on } \Gamma, \tag{20}$$

and

$$\begin{aligned} d^2J(\Omega; V; W) = \int_\Gamma \Bigg\{ &\operatorname{div}\left\{\left[\frac{1}{2}(g - y_d)^2 + \frac{\partial}{\partial n}(y - g)\frac{\partial p}{\partial n}\right] V\right\} \\ &+ \frac{\partial}{\partial n}(y - g)\frac{\partial P_V'}{\partial n} + \frac{\partial p}{\partial n}\frac{\partial Y_V'}{\partial n} \Bigg\} W \bullet n \, d\Gamma, \end{aligned} \tag{21}$$

where we have added the subscript V to P' and Y' to emphasize that they both depend on V. The last step consists in the elimination of P_V', which will introduce Y_W'. To do that, we set $\psi = \hat{\Psi}_V|_{\Omega_t} = P_V'$ in (10) with $V = W$and $t = 0$

$$\begin{aligned} &\int_\Omega \left[\operatorname{div}\{[\nabla(y - g) \bullet \nabla\psi]W\} + \Delta\psi Y_W' + \psi\Delta Y_W' + \nabla\psi \bullet \nabla Y_W'\right] dx = 0 \\ &\Rightarrow \int_\Omega \nabla(y - g) \bullet \nabla P_V' W \bullet n \, d\Gamma + \int_\Omega \Delta P_V' Y_W' + P_V' \Delta Y_W' + \nabla P_V' \bullet \nabla Y_W' \, dx = 0 \end{aligned} \tag{22}$$

and $\phi = \hat{\Phi}_W|_{\Omega_t} = Y_W'$ in (11) with $t = 0$,

$$\begin{aligned} &\int_\Omega \left[\operatorname{div}\{[\nabla\Phi \bullet \nabla p]V\} + \Delta\Phi P_V' + \Phi\Delta P_V' + \nabla\Phi \bullet \nabla P_V' + \Phi Y_V'\right] dx = 0 \\ &\Rightarrow \int_\Gamma \nabla Y_W' \bullet \nabla p \; V \bullet n \, d\Gamma + \int_\Gamma \Delta Y_W' P_V' + Y_W' \Delta P_V' + \nabla Y_W' \bullet \nabla P_V' + Y_W' Y_V' \, dx = 0. \end{aligned} \tag{23}$$

This yields the following identity:

$$\int_\Gamma \nabla(y - g) \bullet \nabla P_V' \; W \bullet n \, d\Gamma = \int_\Gamma \nabla Y_W' \bullet \nabla p \; V \bullet n \, d\Gamma + \int_\Gamma Y_W' Y_V' \, dx \tag{24}$$

or

$$\int_{\Gamma} \frac{\partial}{\partial n}(y-g)\frac{\partial P'_V}{\partial n} W \bullet n \, d\Gamma = \int_{\Gamma} \frac{\partial Y'_W}{\partial n}\frac{\partial p}{\partial n} V \bullet n \, d\Gamma + \int_{\Gamma} Y'_W Y'_V \, dx. \tag{25}$$

As a result

$$\begin{aligned} d^2 J(\Omega; V; W) = \int_{\Gamma} & \operatorname{div}\left\{\left[\frac{1}{2}(g-y_d)^2 + \frac{\partial}{\partial n}(y-g)\frac{\partial p}{\partial n}\right] V\right\} W \bullet n \, d\Gamma \\ & + \int_{\Gamma} \frac{\partial p}{\partial n}\left[\frac{\partial Y'_W}{\partial n} V \bullet n + \frac{\partial Y'_V}{\partial n} W \bullet n\right] d\Gamma + \int_{\Gamma} Y'_W Y'_V \, dx, \end{aligned} \tag{26}$$

where Y'_V is the unique solution of

$$\Delta Y'_V = 0 \quad \text{in } \Omega, \qquad Y'_V = -\frac{\partial}{\partial n}(y-g) V \bullet n \quad \text{on } \Gamma. \tag{27}$$

5.2. Verification of the hypotheses. In §5.1, we have boldly applied the conclusion of the theorem of Correa and Seeger to a Lagrangian that contains a cost functional that is not necessarily convex. This means that the corresponding Lagrangian functional does not necessarily have saddle points. Yet, the conclusions of the theorem extend to *semiconvex* cost functionals (§5.2.1). The verification of the hypotheses will essentially be the same as for the gradient in §4.2 (§5.2.2).

5.2.1. Semiconvex cost functionals. Consider a Lagrangian functional of the form

$$G(t,x,y) = F(t,x) + b(t,x,y) \tag{28}$$

for a family of continuous bilinear forms $b(t,x,y)$ on $X \times Y$ and continuous cost functionals $F(t,x)$ on X. Formally, the saddle point equations are given by

$$x^t \in X, \quad b(t,x^t,y) = 0 \quad \forall y \in Y \tag{29}$$

$$y^t \in Y, \quad dF(t,x^t;x) + b(t,x,y^t) = 0 \quad \forall x \in X. \tag{30}$$

When $G(t,x,y)$ is convex in x and concave in y, (29)–(30) characterize the saddle points $X(t) \times Y(t) \subset X \times Y$ of $G(t,\cdot,\cdot)$. So, when $F(t,x)$ is not convex in x, (29)–(30) need not characterize saddle points of $G(t,\cdot,\cdot)$.

We say that the functional $F(t,x)$ is *semiconvex* in x if there exists a family of continuous convex functionals $C(t,x)$ on X such that $F(t,x) + C(t,x)$ is convex in x. This means that $F(t,\cdot) + C(t,\cdot)$ and $C(t,\cdot)$ both have directional derivatives and hence $F(t,\cdot)$ also has a directional derivative: the following limit exists

$$dF(t,x;x') = \lim_{\theta \searrow 0} \frac{F(t,x+\theta x') - F(t,x)}{\theta}. \tag{31}$$

Denote by $X(t)$ the set of all solutions

$$x^t \in X, \quad b(t,x^t,y) = 0 \quad \forall y \in Y \tag{32}$$

and assume that

$$\forall x^t \in X(t), \quad F(t,x^t) = J(t); \tag{33}$$

that is, $F(t,x^t)$ is only a function of t. We use $J(t)$ as the definition of our cost function.

Now, assume that $F(t,\cdot)$ is semiconvex and that

$$\forall x^t \in X(t), \quad C(t,x^t) = J_0(t); \tag{34}$$

that is $C(t,x^t)$ is only a function of t. Again, use $J_0(t)$ as the definition of the cost function associated with C. Finally, let

$$J_C(t) = F(t,x^t) + C(t,x^t), \tag{35}$$

which is also only a function of t. Then it is obvious that

$$J(t) = J_C(t) - J_0(t) \tag{36}$$

and that (if $dJ_C(0)$ and $dJ_0(0)$ exist)

$$dJ(0) = dJ_C(0) - dJ_0(0). \tag{37}$$

So, we are back to the use of the theorem of Correa and Seeger [1] for both J_C and J_0. Construct the Lagrangians

$$G_C(t,x,y){=}F(t,x) + C(t,x) + b(t,x,y), \tag{38}$$

$$G_0(t,x,y) {=}C(t,x) + b(t,x,y) \tag{39}$$

and assume that if all the hypotheses are verified both for G_0 and G_C,

$$J_0(t) = \operatorname*{Min}_{x\in X} \ \operatorname*{Max}_{y\in Y} \ G_0(t,x,y), \tag{40}$$

$$dJ_0(0) = \operatorname*{Min}_{x\in X(0)} \ \operatorname*{Max}_{y\in Y_0(0)} \ \partial_t G_0(0,x,y), \tag{41}$$

$$J_C(t) = \operatorname*{Min}_{x\in X} \ \operatorname*{Max}_{y\in Y} \ G_C(t,x,y), \tag{42}$$

$$dJ_C(0) = \operatorname*{Min}_{x\in X(0)} \ \operatorname*{Max}_{y\in Y_C(0)} \ \partial_t G_C(0,x,y), \tag{43}$$

where the saddle point equations for J_0 are given by

$$x^t \in X, \quad B(t,x^t,y) = 0 \quad \forall y, \tag{44}$$

$$y_0^t \in Y, \quad dC(t,x^t;x) + b(t,x,y_0^t) = 0 \quad \forall x, \tag{45}$$

and the saddle point equations for $J_C(t)$

$$x^t \in X, \quad B(t,x^t,y) = 0 \quad \forall y, \tag{46}$$

$$y_C^t \in Y, \quad dC(t,x^t;x) + dF(t,x^t;x) + b(t,x,y_C^t) = 0 \quad \forall x. \tag{47}$$

Assume that

$$dJ_0(0) = \partial_t C(0,x^0) + \partial_t b(0,x^0,y_0^0), \ \forall x^0 \in X(0), \ \forall y_0^0 \in Y_0(0), \tag{48}$$

and that $dC(0,x^0;x)$ and $dF(0,x^0;x)$ are independent of the point x^0 chosen in $X(0)$. By subtracting (45) from (47), construct the new variable $y^0 = y_C^0 - y_0^0$, which is a solution of

$$y^0 \in Y, \ dF(0,x^0;x) + b(0,x,y^0) = 0, \ \forall x, \tag{49}$$

and the set

$$Y(0) = \{y_C^0 - y_0^0 : y_C^0 \in Y_C(0), \ y_0^0 \in Y_0(0)\}.$$

Now,

$$\begin{aligned} dJ_C(0) &= \operatorname*{Min}_{x\in X(0)} \operatorname*{Max}_{y\in Y_C(0)} \partial_t F(0,x) + \partial_t C(0,x) + \partial_t b(0,x,y) \\ &= \operatorname*{Min}_{x\in X(0)} \operatorname*{Max}_{\substack{y_0^0\in B_0(0)\\ y_0\in Y(0)}} \{\partial_t F(0,x) + \partial_t b(0,x,y^0) + \partial_t C(0,x) + \partial_t b(0,x,y_0^0)\}. \end{aligned}$$

However, for all $(x^0, y_0^0) \in X(0) \times Y_0(0)$,

$$\partial_t C(0,x^0) + \partial_t b(0,x^0,y_0^0) = dJ_0(0),$$

and finally,

$$dJ(0) = dJ_C(0) - dJ_0(0) = \operatorname*{Min}_{x\in X(0)} \operatorname*{Max}_{y\in Y(0)} [\partial_t F(0,x) + \partial_t b(0,x,y)], \tag{50}$$

where the saddle point $(x^0, y^0) \in X(0) \times Y(0)$ is the solution of the "formal saddle point equations" (29)–(30) for $t = 0$.

In §4.2, the cost functional is semiconvex since there exists a constant $C > 0$ large enough so that

$$dJ(\Omega_t(W); V) + C\ [||y_t||^2_{H^4(\Omega_t)} + ||p_t||^2_{H^4(\Omega_t)}\ ] \tag{51}$$

is convex and continuous on $H^4(\Omega_t) \times H^4(\Omega_t)$. The functional

$$C(t,\phi,\psi) = c||\phi||^2_{H^4(\Omega_t)} + ||\psi||^2_{H^4(\Omega_t)} \tag{52}$$

is clearly convex and continuous on $H^4(\Omega_t) \times H^4(\Omega_t)$. This provides a complete justification for the use of the conclusions of Correa and Seeger.

5.2.2. Verification of the hypotheses. We have chosen to work in $H^4(\mathbb{R}^N) \times H^4(\mathbb{R}^N) \times H^3(\mathbb{R}^N) \times H^3(\mathbb{R}^N)$ and introduced appropriate hypotheses on f, y_d, and g in §5.1. From this point on, the technique is the same as the one used in §4.2 for the gradient. Therefore, we will not repeat it here.

6. Appendix. Proofs of the theorems of §2.

Proof of Theorem 2.1. (i) Properties (T1) follow by standard arguments.

Condition (T2). Associate with X in $\mathbb{R}^N$ the function

$$y(s) = T_{t-s}(X), \qquad 0 \le s \le t. \tag{1}$$

Then

$$\frac{dy}{ds}(s) = -V(t-s, y(s)), \quad 0 \le s \le t, \quad y(0) = T_t(X). \tag{2}$$

For each $x \in \mathbb{R}^N$, the differential equation

$$\frac{dy}{ds}(s) = -V(t-s, y(s)), \quad 0 \le s \le t, \quad y(0) = x \in \mathbb{R}^N \tag{3}$$

has a unique solution in $C^1([0,t]; \mathbb{R}^N)$. The solutions of (3) define a Lipschitzian mapping

$$x \mapsto S_t(x) = y(t) : \mathbb{R}^N \to \mathbb{R}^N \tag{4}$$

such that

$$\exists c > 0, \quad \forall t \in [0,\tau], \quad \forall x, y \in \mathbb{R}^N, \quad |S_t(y) - S_t(x)| \le c|y - x|. \tag{5}$$

Now in view of (2) and (3)

$$S_t(T_t(X)) = y(t) = T_{t-t}(X) = X \Rightarrow S_t \circ T_t = I \quad \text{on } \mathbb{R}^N.$$

To obtain the other identity, consider the function

$$z(r) = y(t-r;x),$$

where $y(\cdot, x)$ is the solution of (3) for some arbitrary x in $\mathbb{R}^N$. By definition

$$\frac{dz}{dr}(r) = V(r, z(r)), \quad z(0) = y(t,x)$$

and necessarily

$$\begin{aligned} x = y(0;x) = z(t) = T_t(y(t;x)) = T_t(S_t(x)) \\ \Rightarrow T_t \circ S_t = I \quad \text{on } \mathbb{R}^N \Rightarrow S_t = T_t^{-1} : \mathbb{R}^N \to \mathbb{R}^N. \end{aligned}$$

Condition (T3). The uniform Lipschitz continuity in (T3) follows from (5) and we only need to show that

$$\forall x \in \mathbb{R}^N, \quad T^{-1}(\cdot, x) \in C^0([0,\tau]; \mathbb{R}^N).$$

Given t in $[0,\tau]$, pick an arbitrary sequence $\{t_n\}, t_n \to t$. Then for each $x \in \mathbb{R}^N$ there exists $X \in \mathbb{R}^N$ such that

$$T_t(X) = x \quad \text{and} \quad T_{t_n}(X) \to T_t(X) = x.$$

But

$$T_{t_n}^{-1}(x) - T_t^{-1}(x) = T_{t_n}^{-1}(T_t(X)) - T_t^{-1}(T_t(X)) = T_{t_n}^{-1}(T_t(X)) - T_{t_n}^{-1}(T_{t_n}(X)).$$

By the uniform Lipschitz continuity of T_t^{-1},

$$|T_{t_n}^{-1}(x) - T_t^{-1}(x)| = |T_{t_n}^{-1}(T_t(X)) - T_{t_n}^{-1}(T_{t_n}(X))| \le c|T_t(X) - T_{t_n}(X)|$$

and the last term converges to zero as t_n goes to t.

(ii) The first part of condition (V) is verified since for each $x \in \mathbb{R}^n$ and t, s in $[0,\tau]$,

$$\begin{aligned} &|V(t,x) - V(s,x)| \\ &\le \left|\frac{\partial T}{\partial t}(t, T_t^{-1}(x)) - \frac{\partial T}{\partial t}(t, T_s^{-1}(x))\right| + \left|\frac{\partial T}{\partial t}(t, T_s^{-1}(x)) - \frac{\partial T}{\partial t}(s, T_s^{-1}(x))\right| \\ &\le c|T_t^{-1}(x) - T_s^{-1}(x)| + \left|\frac{\partial T}{\partial t}(t, T_s^{-1}(x)) - \frac{\partial T}{\partial t}(s, T_s^{-1}(x))\right|. \end{aligned}$$

So from (T3) and (T1) $s \mapsto V(s,x)$ is continuous at $s = t$ and hence for all x in $\mathbb{R}^N$, $V(\cdot, x) \in C^0([0,\tau]; \mathbb{R}^N)$. The Lipschitzian property follows directly from the Lipschitzian properties (T1) and (T3): for all x and y in $\mathbb{R}^N$,

$$\begin{aligned} |V(t,y) - V(t,x)| &= \left|\frac{\partial T}{\partial t}(t, T_t^{-1}(y)) - \frac{\partial T}{\partial t}(t, T_t^{-1}(x))\right| \\ &\le c\left|T_t^{-1}(y) - T_t^{-1}(x)\right| \le cc'|y-x|. \end{aligned}$$

This proves that V verifies condition (V). □

Proof of Theorem 2.2. (i) By definition of T in (2.8), $t \mapsto T(t,X)$ and $t \mapsto (\partial T/\partial t)(t,X) = U(X) + tA(X)$ are continuous on $[0,\infty[$. Moreover, for all X and Y,

$$|T(t,Y) - T(t,X)| \le \left[1 + tc + \frac{t^2}{2}c\right] |Y - X|$$

and

$$\left|\frac{\partial T}{\partial t}(t,Y)-\frac{\partial T}{\partial t}(t,X)\right| \leq \ [c+tc]\ |Y-X|.$$

Thus condition (T1) is verified for any finite $\tau > 0$. To check condition (T2) we need to prove that $X \mapsto T_t(X) = T(t,X) : \mathbb{R}^N \mapsto \mathbb{R}^N$ is bijective. For all Y and X,

$$T_t(Y)-T_t(X) = Y - X + t[U(Y)-U(X)] + \frac{t^2}{2}[A(Y)-A(X)]$$

and

$$|T_t(Y)-T_t(X)| \geq |Y-X| - ct|Y-X| - c\frac{t^2}{2}|Y-X|.$$

So for $\tau = \min\{1, 1/4c\}$ and for all t in $[0,\tau]$

$$(6) \qquad |T_t(Y)-T_t(X)| \geq \frac{1}{2}|Y-X|.$$

Hence T_t is injective and, a fortiori, bijective since $\mathbb{R}^N$ is finite-dimensional.

The last part of the proof is the uniform Lipschitzian property of T_t^{-1}. In view of (6) for all x and y

$$(7) \qquad \frac{1}{2}|T_t^{-1}(y)-T_t^{-1}(x)| \ \leq \ |T_t(T_t^{-1}(y))-T_t(T_t^{-1}(x))| \ = \ |y-x|.$$

To complete our argument, we prove the continuity with respect to t for each x. Let $X = T_t^{-1}(x)$. For any s in $[0,\tau]$

$$T_s^{-1}(x)-T_t^{-1}(x) \ = \ T_s^{-1}(T_t(X))-T_t^{-1}(T_t(X)) \ = \ T_s^{-1}(T_t(X))-T_s^{-1}(T_s(X))$$

and in view of (7)

$$|T_s^{-1}(x)-T_t^{-1}(x)| \ \leq \ 2|T_t(X)-T_s(X)|.$$

The continuity of $T_s^{-1}(x)$ at $s=t$ now follows from the continuity of $T_s(X)$ at $s=t$. Thus condition (T3) is verified. □

Proof of Theorem 2.3. By definition of the inductive limit topology on $\overrightarrow{\mathcal{V}}^{m,k}$, it is sufficient to prove the continuity for any compact subset K of $\mathbb{R}^N$. So given V in $\mathcal{V}_K^{m,k}$ construct the sequence

$$V_n(t) = V(t/n), \quad 0 \leq t \leq \tau, \quad \text{for integers } n \geq 1.$$

By continuity of V, $\{V_n\}$ converges in $V_K^{m,k}$ to the autonomous field $\tilde{V}$, $\tilde{V}(t,x) = V(0,x)$ for all $t \in [0,\tau]$. Hence by continuity of (2.17)

$$dJ(\Omega;V_n) \rightarrow dJ(\Omega;V(0))$$

and by uniqueness of the limit we obtain (2.18). □

Proof of Theorem 2.4. (i) For any V in $\mathcal{D}^k(\mathbb{R}^N,\mathbb{R}^N)$ such that $V=0$ on Γ, we have by Nagumo's [1] theorem (cf. also Aubin and Cellina [1] for an English version)

$$\forall t \in [0,\tau], \quad T_t(\bar{\Omega}) \subset \bar{\Omega}.$$

Now always by Nagumo's [1] theorem applied to (3) in the proof of Theorem 2.1

$$\forall t \in [0,\tau], \quad T_t^{-1}(\bar{\Omega}) = S_t(\bar{\Omega}) \subset \bar{\Omega}.$$

Necessarily, $T_t(\bar{\Omega}) = \bar{\Omega}$ and since T_t is a homeomorphism and Ω is open

$$\forall t \in [0,\tau], \quad T_t(\Omega) = \Omega \quad \Rightarrow \quad \forall t \in [0,\tau], \quad J(\Omega_t) = J(\Omega) \quad \Rightarrow \quad dJ(\Omega;V) = 0.$$

(ii) Since the open domain Ω has a boundary Γ that is C^{k+1}, there exists a unique outward normal $n(x)$ that belongs to $C^k(\Gamma, \mathbb{R}^N)$. Define the subspace

$$L^k_\Omega \ = \ \{V \in \mathcal{D}^k(\mathbb{R}^N, \mathbb{R}^N) \ : \ (\gamma_\Gamma V) \bullet n \ = \ 0 \text{ on } \Gamma\}$$

of $\mathcal{D}^k(\mathbb{R}^N, \mathbb{R}^N)$. It is closed and linear. Moreover, by using Nagumo's [1] theorem twice for all V in L^k_Ω, the corresponding transformation T_t has the property

$$T_t(\Omega) = \Omega, \quad \forall t \in [0, \tau]$$

and necessarily

$$dJ(\Omega; V) \ = \ 0 \quad \Rightarrow \quad V \in \ker(G(\Omega)).$$

This induces a continuous linear map

$$[G(\Omega)] \ : \ \mathcal{D}^k(\mathbb{R}^N, \mathbb{R}^N)/L^k_\Omega \ \to \ \mathbb{R} \tag{8}$$

such that

$$\langle [G(\Omega)], q_L(V)\rangle_{\mathcal{D}^k/L^k} \ = \ dJ(\Omega; V) \ = \ \langle G(\Omega), V\rangle_{\mathcal{D}^k},$$

where

$$q_L \ : \mathcal{D}^k(\mathbb{R}^N, \mathbb{R}^N) \ \to \ \mathcal{D}^k(\mathbb{R}^N, \mathbb{R}^N)/L^k_\Omega \tag{9}$$

is the canonical surjection. For a boundary Γ that is C^{k+1}, there exists a unique outward normal $n(x)$ that belongs to $C^k(\Gamma, \mathbb{R}^N)$. As a result, the kernel of the map

$$V \ \mapsto \ \gamma_\Gamma V \bullet n \ : \ \mathcal{D}^k(\mathbb{R}^N, \mathbb{R}^N) \ \to \ \mathcal{D}^k(\Gamma) \tag{10}$$

coincides with L^k_Ω, where γ_Γ is the trace operator from $\mathbb{R}^N$ to Γ. Moreover, the map (10) is surjective since it is always possible for a C^{k+1} boundary ($k \geq 0$) to construct a C^k-extension N on $\mathbb{R}^N$ of the unit normal n on Γ (cf. Agmon, Douglis, and Nirenberg [1], [2]). Similarly, for any v in $\mathcal{D}^k(\Gamma)$, there exists an extension $\tilde{v}$ in $\mathcal{D}^k(\mathbb{R}^N)$ and the vector $V = \tilde{v}N$ belongs to $\mathcal{D}^k(\mathbb{R}^N, \mathbb{R}^N)$ and coincides with vn on Γ. As a result, the map

$$q_L(V) \ \mapsto \ p_L(q_L(V)) \ : \ \mathcal{D}(\mathbb{R}^N, \mathbb{R}^N)/L^k_\Omega \ \to \ \mathcal{D}(\Gamma) \tag{11}$$

is a well-defined isomorphism. In particular, there exists a scalar distribution $g(\Omega)$ in $\mathcal{D}(\Gamma)'$ such that (2.20) is verified. □

Proof of Theorem 2.5. The differential quotient (2.24) can be split into the sum of two terms

$$[dJ(\Omega_t(W); V(0)) - dJ(\Omega; V(0))]/t + [dJ(\Omega_t(W); V(t)) - dJ(\Omega_t(W); V(0))]/t. \tag{12}$$

In view of Theorem 2.5 (i) and (iii), for all U in $\mathcal{V}^\ell$

$$d^2J(\Omega; U; W) = d^2J(\Omega; U; W(0))$$

by the same argument as in the proof of Theorem 2.4 for the gradient. Hence the first term converges to

$$d^2J(\Omega; V(0); W) = d^2J(\Omega; V(0); W(0)).$$

For the second term recall that V belongs to $\overset{\to m+1,\ell}{\mathcal{V}}$ and observe that the vector field

$$\tilde{V}(t) = [V(t) - V(0)]/t$$

belongs to $\overset{\to m,\ell}{\mathcal{V}}$ and that $\tilde{V}(0) = \overset{\bullet}{V}(0)$. Thus by linearity of $dJ(\Omega; V)$ the second term in (12) can be written as

$$dJ(\Omega_t(W); [V(t) - V(0)]/t) = dJ(\Omega_t(W); \tilde{V}(t)).$$

For any V in $\overset{\to m+2,\ell}{\mathcal{V}}$, $\tilde{V}$ belongs to $\overset{\to m+1,\ell}{\mathcal{V}}$. Then by assumption (i)

$$\lim_{t\searrow 0}[dJ(\Omega_t(W);\tilde{V}(t)) - dJ(\Omega;\tilde{V}(0))]/t = d^2J(\Omega;\tilde{V};W),$$

which implies that

$$\lim_{t\searrow 0} dJ(\Omega_t(W);\tilde{V}(t)) = dJ(\Omega;\tilde{V}(0)) = dJ(\Omega;\overset{\bullet}{V}(0)).$$

Now by hypothesis (ii), $U \mapsto dJ(\Omega;U)$ is linear and continuous on $\mathcal{D}^{\ell}(\mathbb{R}^N,\mathbb{R}^N)$ and the map

$$V \mapsto \overset{\bullet}{V}(0) \mapsto dJ(\Omega;\overset{\bullet}{V}(0)) : \overset{\to m+2,\ell}{\mathcal{V}} \to \mathcal{V}^{\ell} \to \mathbb{R}$$

is linear and continuous (hence uniformly continuous) for the topology $\overset{\to m+1,\ell}{\mathcal{V}}$ for all V in the dense subspace $\overset{\to m+2,\ell}{\mathcal{V}}$. Hence it uniquely and continuously extends to all elements of $\overset{\to m+1,\ell}{\mathcal{V}}$. This completes the proof of the theorem. □

Proof of Theorem 2.6. (i) It is sufficient to prove that

(a) for all $V, W \in \mathcal{D}(\mathbb{R}^N,\mathbb{R}^N)$ such that $W = 0$ in a neighbourhood of Γ, $d^2J(\Omega;V;W) = 0$ and

(b) for all $V, W \in \mathcal{D}(\mathbb{R}^N,\mathbb{R}^N)$ such that $V = 0$ in a neighbourhood of Γ, $d^2J(\Omega;V;W) = 0$.

In case (a) the proof is similar to the one in Theorem 2.4 for the gradient, and we even have the stronger result that for W such that $W = 0$ on Γ

$$\Omega_t(W) = \Omega, \quad \forall t \in [0,\tau] \quad \Rightarrow \quad dJ(\Omega_t(W);V) = dJ(\Omega;V) \quad \Rightarrow \quad d^2J(\Omega;V;W) = 0.$$

In case (b) $V = 0$ in a neighbourhood N of Γ and in $\mathbb{R}^N\backslash K$, the complement of the compact supports K of V. So $U = (\mathbb{R}^N\backslash K)$ is a neighbourhood of Γ where $V = 0$. By construction $U \cap K = \emptyset$ and there exists a bounded neighbourhood $\mathcal{U}$ of K such that $\bar{\mathcal{U}} \cap \Gamma = \emptyset$. Since $\bar{\mathcal{U}}$ is compact and Γ is closed, the minimum distance d from $\bar{\mathcal{U}}$ to Γ is finite and nonzero. Let

$$N(\Gamma) = \{y \in \mathbb{R}^N : d_\Gamma(y) < d/2\},$$

where

$$d_\Gamma(y) = \inf\{|y-x| : x \in \Gamma\}.$$

For all X in Γ

$$T_t(X) - X = \int_0^t W(T_s(X))\,ds = tW(X) + \int_0^t [W(T_s(X)) - W(X)]\,ds$$

and since W is uniformly Lipschitzian by assumption (V) on W

$$|T_t(X) - X| \le t|W(X)| + ct \max_{[0,t]} |T_s(X) - X|.$$

It can easily be shown that for $t < 1/c$

$$\max_{[0,t]} |T_s(X) - X| < \frac{t}{1-ct}|W(X)|.$$

Thus

$$\sup_{X\in\Gamma} \max_{[0,t]} |T_s(X) - X| \le \frac{t}{1-ct} \sup_{X\in\Gamma} |W(X)|.$$

However, W is continuous with compact support. Therefore,

$$\sup_{X\in\Gamma}|W(X)| \le \sup_{X\in \operatorname{supp} W}|W(X)| = \|W\|_{C^0(\mathbb{R}^N;\mathbb{R}^N)} < \infty$$

and there exists $\tau > 0$ such that

$$\forall s \in [0,\tau], \quad \frac{s}{1-cs}\|W\|_{C^0} < \frac{d}{2}.$$

By definition and the previous inequalities

$$d_\Gamma(T_s(X)) = \inf_{Y\in\Gamma}|T_s(X)-Y| \le |T_s(X)-X| < \frac{d}{2}$$

for all s in $[0,\tau]$ and all $X\in\Gamma$. This implies that

$$\forall s\in[0,\tau], \forall X\in\Gamma, \quad \Gamma_s(W) = T_s(W)(\Gamma) \subset N(\Gamma).$$

By construction $V = 0$ in $N(\Gamma)$ since the distance from K to Γ is greater than or equal to d. Therefore,

$$\forall s\in[0,\tau], \quad V \in L^\infty_{\Omega_s(W)}$$

and as in the proof of Theorem 2.4(ii), for all $s > 0$ $dJ(\Omega_s(W);V) = 0$ and necessarily $d^2J(\Omega;V;W) = 0$.

(ii) We have already established in (i) that the bilinear form

$$(V,W) \mapsto h(V,W) : \mathcal{D}(\mathbb{R}^N,\mathbb{R}^N)\times\mathcal{D}(\mathbb{R}^N,\mathbb{R}^N) \to \mathbb{R}$$

is zero for all $V\in\mathcal{D}(\mathbb{R}^N,\mathbb{R}^N)$ and $W\in\mathcal{D}(\mathbb{R}^N,\mathbb{R}^N)$ such that $W=0$ on Γ and also zero for all $W\in\mathcal{D}(\mathbb{R}^N,\mathbb{R}^N)$ and $V\in\mathcal{D}(\mathbb{R}^N,\mathbb{R}^N)$ for which $V=0$ in a neighbourhood of Γ. By density all this is still true in $\mathcal{D}^\ell(\mathbb{R}^N,\mathbb{R}^N)$ and now by the same argument as in the proof of Theorem 2.4 for all V in $\mathcal{D}^\ell(\mathbb{R}^N,\mathbb{R}^N)$

$$[W] \mapsto h(V,W) : \mathcal{D}^\ell(\mathbb{R}^N,\mathbb{R}^N)/L^\ell_\Omega \to \mathbb{R}$$

is well defined, linear, and continuous. For the first component it is necessary to show that for all W in

$$D^\ell_\Gamma = \{V\in\mathcal{D}^\ell(\mathbb{R}^N,\mathbb{R}^N) : \partial^\alpha V = 0 \text{ on } \Gamma,\ \forall|\alpha|\le\ell\},$$

the bilinear form $h(V,W)=0$. We first prove the result for the subspace

$$A = \mathcal{D}(\Omega;\mathbb{R}^N)\oplus\mathcal{D}(\mathbb{R}^N\backslash\bar\Omega;\mathbb{R}^N).$$

Then we prove that $\bar A = D^\ell_\Gamma$. Finally, by density and continuity, the result holds for the $\mathcal{D}^\ell(\mathbb{R}^N,\mathbb{R}^N)$-closure $\bar A$ of A.

For any V in A, there exist $V_1\in\mathcal{D}(\Omega;\mathbb{R}^N)$ and $V_2\in\mathcal{D}(\mathbb{R}^N\backslash\bar\Omega;\mathbb{R}^N)$ such that $V = V_1+V_2$. Moreover,

$$K_1 = \operatorname{supp} V_1 \subset \Omega \quad\text{and}\quad K_2 = \operatorname{supp} V_2 \subset \mathbb{R}^N\backslash\bar\Omega$$

are compact subsets of the open sets Ω and $\mathbb{R}^N\backslash\bar\Omega$, respectively. Hence $V_1 = 0$ (respectively, $V_2=0$) in the open neighbourhood $\mathbb{R}^N\backslash K_1$ (respectively, $\mathbb{R}^N\backslash K_2$) of Γ and necessarily $V = V_1+V_2 = 0$ in the neighbourhood $U = \mathbb{R}^N\backslash(K_1\cup K_2)$ of Γ. Hence from part (i) $h(V,W)=0$.

By definition of D^ℓ_Γ,

$$D^\ell_\Gamma \subset \mathcal{D}^\ell(\bar\Omega;\mathbb{R}^N)\oplus\mathcal{D}^\ell(\mathbb{R}^N\backslash\Omega;\mathbb{R}^N).$$

Now

$$A \subset D^\ell_\Gamma \text{ and } \bar{A} = \overline{\mathcal{D}(\Omega;\mathbb{R}^N)} \oplus \overline{\mathcal{D}(\mathbb{R}^N\backslash\bar{\Omega};\mathbb{R}^N)} \subset D^\ell_\Gamma$$

and

$$\overline{\mathcal{D}(\Omega;\mathbb{R}^N)} = \mathcal{D}^\ell(\bar{\Omega};\mathbb{R}^N) \quad \text{and} \quad \overline{\mathcal{D}(\mathbb{R}^N\backslash\bar{\Omega};\mathbb{R}^N)} = \mathcal{D}^\ell(\mathbb{R}^N\backslash\Omega;\mathbb{R}^N).$$

This proves that $\bar{A} = D^\ell_\Gamma$. To complete the proof, note, that by continuity of $V \mapsto h(V,W)$, for all W in $\mathcal{D}^\ell(\mathbb{R}^N,\mathbb{R}^N)$ the map

$$[V] \mapsto h(V,W) : \mathcal{D}^\ell(\mathbb{R}^N,\mathbb{R}^N)/D^\ell_\Gamma \to \mathbb{R}$$

is well defined, linear, and continuous. Finally, the map

$$([V],[W]) \mapsto h(V,W) : (\mathcal{D}^\ell(\mathbb{R}^N,\mathbb{R}^N)/L^\ell_\Omega) \times (\mathcal{D}^\ell(\mathbb{R}^N,\mathbb{R}^N)/D^\ell_\Gamma) \to \mathbb{R}$$

is well defined, bilinear, and continuous. □

REFERENCES

S. AGMON, A. DOUGLIS, AND L. NIRENBERG [1], *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions,* I, Comm. Pure Appl. Math. 12 (1959), 623–727.

[2], *Estimates near the boundary for solutions of elliptic partial differential equations satisfying general boundary conditions*, II, Comm. Pure Appl. Math. 17 (1964), 35–92.

G. ARUMUGAM AND O. PIRONNEAU [1], *On the problems of riblets as a drag reduction device*, Optimal Control Appl. Methods 10 (1989), 93–112.

[2], *Sur le problème des "riblets,* " Rapport de Recherche R87027, Publications du Laboratorei d'Analyse Numérique, Université Pierre et Marie Curie, Paris, France, 1987.

J. P. AUBIN, AND A. CELLINA [1], "*Differential Inclusions*, Springer-Verlag, Berlin, 1984.

V. M. BABIĆ [1], *Sur le prolongement des fonctions in Russian*, Uspekhi Mat. Nauk 8 (1953), 111–113.

A. BERN [1], "*Thèse de l'École Nationale Supérieure des Mines de Paris*, CEMEF, Sophia Antipolis, France, October 1987.

A. BERN, J. L. CHENOT, Y. DEMAY, AND J.-P. ZOLÉSIO [1], *Numerical computation of the free boundary in non-Newtonian stationary flows*, Proc. 6th Internat. Symposium on Finite Element Methods in Flow Problems, June 1986, 383–390, Publications INRIA, Rocquencourt, France.

P. CANNARSA AND H. M. SONER [1], *On the singularities of the viscosity solutions to Hamilton–Jacobi–Bellman Equations*, Indiana University Math. J. 36 (1987), 501–524.

J. CÉA [1], *Problems of shape optimal design*, in "Optimization of Distributed Parameter Structures, vol II, E. J. Haug and J. Céa, eds., Sijhtoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981, pp. 1005–1048.

[2], *Numerical methods of shape optimal design*, in "Optimization of Distributed Parameter Structures, vol II, E. J. Haug and J. Céa, eds., Sijhoff and Noordhoff, Alphen aan den Rijn, the Netherlands, 1981, pp. 1049–1087.

[3], *Conception optimale ou identification de formes*: calcul rapide de la dérivée directionnelle de la fonction coût, Math. Modelling Numer. Anal. 20 (1986), 371–402.

K.-T. CHENG AND N. OLHOFF [1], *Regularized formulation for optimal design of axisymmetric plates*, Internat. J. Solids and Structures 18 (1982), 153–169.

R. CORREA AND A. SEEGER [1], *Directional derivatives of a minimax function*, Nonlinear Anal. 9 (1985), 13–22.

M. C. DELFOUR, G. PAYRE, AND J.-P. ZOLÉSIO [1], *Shape optimal design of a radiating fin*, in "System Modelling and Optimization, P. Thoft-Christensen, ed., Springer-Verlag, Berlin, Heidelberg, 1984, pp. 810–818.

[2], *An optimal triangulation for second order elliptic problems*, in "in Computer Methods in Applied Mechanics and Engineering, 50 (1985), pp. 231–261.

M. C. DELFOUR AND J.-P. ZOLÉSIO [1], *Dérivation d'un MinMax et application à la dérivation par rapport au contrôle d'une observationnon-différentiable de l'état*, C.R. Acad. Sci. Paris t. 302, Sér. I, 16 (1986), 571–574.

[2], *Shape sensitivity analysis via MinMax differentiability*, SIAM J. Control Optim. 26 (1988), 834–862.

[3], *Differentiability of a MinMax and application to optimal control and design problems, Part* I, in "in Control Problems for Systems Described as Partial Differential Equations and Applications, I. Lasiecka and R. Triggiani, eds., Springer-Verlag, New York, 1987, pp. 204–219.

[4], *Differentiability of a MinMax and application to optimal control and design problem, Part* II, in "in Control Problems for Systems Described as Partial Differential Equations and Applications, I. Lasiecka and R. Triggiani, eds., Springer-Verlag, New York, 1987, pp. 220-229.

[5], *Further developments in shape sensitivity analysis via a penalization method*, in "in Boundary Control and Boundary Variations, J.-P. Zolésio, ed., Springer-Verlag, Berlin, Heidelberg, New York, Tokyo, 1988, pp. 153–191.

[6], *Shape sensitivity analysis via a penalization method*, Ann. di Mat. Pura Appl., in CLI (1988), 179–212.

[7], *in Analyse des problèmes de forme par la dérivation des Min Max*, in "in Analyse Non Linéaire, H. Attouch, J. P. Aubin, F. H. Clarke, and I. Ekeland, eds., Série Analyse Non Linéaire, Annales de l'Institut Henri-Poincaré, Special volume in honor of J.-J. Moreau, Gauthier-Villars, Bordas, Paris, France, 1989, pp. 211–228.

[8], *Anatomy of the shape Hessian*, Ann. Mat. Pura Appl. vol. CLVIII (1989) in press.

[9], *Computation of the shape Hessian by a Lagrangian method*, in "in Proc. 5th Symposium on Control of Distributed Parameter Systems, A. El Jai and M. Amouroux, eds., Pergamon Press (to appear), pp. 85–90.

I. EKELAND AND R. TEMAM [1], "*Analyse convexe et problèmes variationnels,*, Gauthier–Villars Dunod, Paris, 1974.

N. FUJII [1], *Domain optimization problems with a boundary value problem as a constraint*, in "in Control of Distributed Parameter Systems 1986, H.E. Rauch, ed., Pergamon Press, Oxford, New York, 1986, pp. 5–9.

[2], *Second variation and its application in a domain optimization problem*, in "in Control of Distributed Parameter Systems 1986, M. Amouroux and A. El Jai, eds., Pergamon Press, Oxford, New York, 1986, pp. 431–436.

M. GUELFAND AND N. Y. VILENKIN [1], "*Les distributions*, *Applications de l'analyse harmonique*, (translated by G. Rideau), Dunod, Paris, 1967.

J. HADAMARD [1], *Mémoire sur le problème d'analyse relatif à l'équilibre des plaques élastiques encastrées*, in "in Œuvres de J. Hadamard, vol. II, CNRS, Paris, 1968, original reference: Mem. Sav. Etrang. 33 (1907), mémoire couronné par l'Académie des Sciences, pp. 515–641.

M. NAGUMO [1], *Über die Loge der Integralkurven gewöhnlicher Differentialgleichungen*, Proc. Phys. Math. Soc. Japan 24 (1942), 551–559.

J. NEČAS [1], "*Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, et Academia, Prague, 1967.

O. PIRONNEAU [1], "*Optimal Design for Elliptic Systems*, Springer-Verlag, New York, 1984.

L. SCHWARTZ [1], "*Théorie des distributions*, Hermann, Paris, 1966.

[2], *Théorie des noyaux*, in "in Proc. of the International Congress of Mathematicians, Vol. 1, 1950, pp. 220–230.

J. SIMON [1], *Second variations for domain optimization problems*, "in Control of Distributed Parameter Systems, Proc. 4th Internat. Conference in Vorau, Birkhäuser Verlag, Basel, July 1988, to appear.

J. P. ZOLÉSIO [1], "*Identification de domaines par déformation*, Thèse de doctorat-d'état, Université de Nice, France, 1979.

[2], *The material derivative (or speed) method for shape optimization*, in "in Optimization of Distributed Parameter Structures, vol. II, E.J. Haug and J. Céa, eds., Sijhtofff and Nordhoff, Alphen aan den Rijn, 1981, pp. 1089–1151.

SIAM J. CONTROL AND OPTIMIZATION.
Vol. 29, No. 6, pp. 1443–1475, September 1991

NUMERICAL METHODS FOR STOCHASTIC SINGULAR CONTROL PROBLEMS*

HAROLD J. KUSHNER† AND LUIZ FELIPE MARTINS‡

Abstract. The paper develops a powerful class of numerical methods for stochastic singular control problems. The basic models used are diffusion or reflected diffusions, but the method is of general applicability. The central idea is that of the Markov chain approximation method, where an approximation to the control problem is found for which an optimal solution is computable, and which is an arbitrarily good approximation to the original problem and its optimal value function. The methods are convenient to program and use (and they have been used with success), and they cover a wide variety of problems. In fact, for the singular problem, they seem to be the only ones currently available. Owing to problems in proving tightness of certain processes that occur in the convergence proofs, the methods of proof used for the nonsingular problems need modifications. Examples of useful approximations, the algorithms, and the convergence proofs are given. To illustrate the power of the methods, two classes of problems are dealt with: the first is a class of discounted problems, and the second is an average-cost-per-unit time problem subject to some constraints, which arises in the study of multicustomer class queueing networks under conditions of heavy traffic. The method is applicable to the more standard singular control and ergodic problems with greater ease.

Key words. numerical methods for stochastic control, singular stochastic control, ergodic stochastic control, Markov chain approximations, weak convergence methods, constrained ergodic problems, optimal stochastic control

AMS(MOS) subject classifications. 93E25, 93E20

1. Introduction. Singular stochastic control problems occur in many forms. They have been the subjects of general studies and have arisen in models of financial economics, minimum-fuel-type problems [1]–[6], the modelling of controlled queuing, and production systems under conditions of heavy traffic [7]–[9], among other places. To date, there are few analytical solutions and (to the authors' knowledge) there is no work on the systematic design and convergence proofs for algorithms for numerical methods for the calculation or approximation of the optimal value function, although numerical calculations (using Markov chain approximations) have been carried out in [13]. In this paper, we develop a class of convergent numerical methods for this computation, which works under quite broad conditions. The methods are extensions of the so-called Markov chain approximation approach of [10], [11]. They are convenient to program and use in applications, and have been used with success by the authors. At this time, they seem to be the only available methods for these problems. Reference [11] is an updated version of [10]. Rather than take a very general approach, the versatility and power of our methods will be illustrated by treating two different classes of problems. The methods used are the same as would be used for other singularly controlled problems. The first problem is a reflected singularly controlled diffusion of the type that arises in the modelling of queues and production systems under heavy traffic conditions [7]. We start with a two-dimensional version, just for notational simplicity. This particular problem has a rich enough structure so

* Received by the editors March 26, 1990; accepted for publication (in revised form) October 20, 1990.

† Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912. The work of this author was supported in part by Air Force Office of Scientific Research grant 89-0015, Army Research Office grant DAAL-03-86-0171, and National Science Foundation grant ECS 89-13351.

‡ Division of Applied Mathematics, Brown University, Providence, Rhode Island 02912, and the Department of Mathematics, Federal University of Rio de Janeiro, Rua Sao Januario 561, Rio de Janeiro 20921, Brazil. The work of this author was supported in part by National Science Foundation grant ECS 89-13351.

that it can be used to illustrate the numerical method for the general singular control problem, as will be seen. Of course, as with any numerical method for solving partial differential equations (PDEs) or related systems, there is a "curse of dimensionality." The algorithms have been successfully used on numerous two- and three-dimensional problems. Our problem differs from those in [1]–[6], mainly in the nature of the boundary. For any numerical method, the state space needs to be bounded, and some condition (reflecting or otherwise) put on the boundary.

The second type of situation with which we work is an ergodic and constrained singular control problem, which has also arisen in heavy traffic analysis [9], [13]. Owing to the constraint, it is more complex than the standard ergodic problem, but the approximation methods and results can be readily adapted to other types of singular control ergodic problems.

Let $D[0,\infty)$ denote the space of functions that are right continuous and have left-hand limits, and with the Skorokhod topology [14], [15]. We use the arrow $\Rightarrow$ to denote weak convergence. All the weak convergence analysis will be on this space or its k-fold products $D^k[0,\infty)$ for appropriate k.

In §2 we define the first class of singular control problems. Section 3 deals with the numerical approximation, states the numerical algorithm, and makes a plausability argument concerning the convergence. The general technique of approximation and numerical solution introduced there is of general use for singular control problems. Section 4 concerns a rescaling that is necessary to deal with the tightness problems in the convergence proof. The actual convergence theorem is proved in §5. To illustrate the applicability of the method to a wide class of singular control problems, a somewhat harder case is discussed in §6, and the trivial adaptations required of the method are given. Section 7 formulates the second class of problems, a singular control problem with an average-cost-per-unit time criterion and with some additional constraints. We chose this class due to the complications and challenges that it presents. However, as will be seen, the numerical method works well and is widely applicable. In §8 the numerical approximation is presented together with several simplifications. The actual algorithm is given in §9. Due to the constraints, a linear programming rather than a dynamic programming method needs to be used. However, if the constraints were not present, any of the dynamic programming-based numerical methods that are usable for ergodic problems for Markov chains can be used. The proof of convergence is in §10. Some numerical data is also presented.

2. A reflected singularly controlled diffusion.

2.1. Definition of the problem. Let $G^i > 0$ and define the two-dimensional box $G = [0, G^1] \times [0, G^2]$. Let P be a degenerate Markov transition matrix, whose spectral radius is less than unity. For any two-dimensional vector X, we write its components as (X^1, X^2). Models of the type (2.1) (with or without controls) arise as heavy traffic limits of certain controlled queueing systems [7], [8], [16]:

$$\begin{aligned} X(t) = X(0) &+ \int_0^t b(X(s))ds + \int_0^t \sigma(X(s))dW(s) \\ &+ J(t) + (I - P')Y(t) - U(t), \quad X(t) \in G \quad \text{for all } t \geq 0. \end{aligned} \tag{2.1}$$

X is two-dimensional, $W(\cdot)$ is a standard R^2-valued Wiener process, the components of $J(\cdot)$, $Y(\cdot)$, and $U(\cdot)$ are in $D[0,\infty)$, and $J(0) = Y(0) = U(0) = 0$. The reflection terms $Y^i(\cdot)$ and $U^i(\cdot)$ are nondecreasing and can increase only when $X^i(t) = 0$ (respectively, $X^i(t) = G^i$).

The term $J(\cdot) = (J^1(\cdot), J^2(\cdot))$ is a "singular" control and is represented as follows: There is $K(\cdot) = (K^{12}(\cdot), K^{21}(\cdot))$ right continuous with $K^{ij}(0) = 0$ and $K^{ij}(\cdot)$ nondecreasing such that

$$J^1 = -K^{12} + c_2K^{21}, \qquad J^2 = c_1K^{12} - K^{21}, \tag{2.2}$$

where $1 \geq c_i > 0$. Define the vectors $v_1 = (-1, c_1)$ and $v_2 = (c_2, -1)$. $K(\cdot)$ is called the control. Since we are concerned only with the numerical method in this paper, the reader is referred to the references for the motivation for the model. We mention only that in [7] the $X(t)$ are the (weak) limits of a scaled queueing system with two processors. The p_{ij} are the probabilities that a completed service at processor i is sent to processor j $(i, j = 1, 2)$, and the K^{ij} represent the weak limit of the suitably scaled number of customers who were intended for processor i but who were actually rerouted to processor j. Without loss of generality, we let $p_{ii} = 0$. Familiarity with [7] is not needed here. As mentioned in the Introduction, the model is canonical in the sense that it has a rich enough structure so that (as will be seen below) the methods are applicable to typical singularly controlled systems of any dimension, and it is chosen for illustrative purposes and to simplify the development only.

It is possible that there will be an impulsive control action at $t = 0$. Then $X(\cdot)$, $J(\cdot)$ will be discontinuous there and will not be right continuous at that point.

2.2. Admissible controls. We say that a control $K(\cdot)$ is *admissible* or that $(K(\cdot), W(\cdot))$ is an *admissible pair* if $K(\cdot)$ is nonanticipative with respect to $W(\cdot)$.

In the absence of control, $Y(\cdot)$ and $U(\cdot)$ are continuous [16], [7]. In the presence of control, they might be discontinuous. For example, and loosely speaking, it is possible that on some time interval, $X(\cdot)$ will be on the boundary of G and the increments of $J(\cdot)$ will "point out of G." Of course, $X(\cdot)$ cannot leave G due to the reflection terms $Y(\cdot)$ and $U(\cdot)$, but such $J(\cdot)$ might cause $Y(\cdot)$ and $U(\cdot)$ to be discontinuous. For the type of problem that motivated the particular example, if G is chosen appropriately, then we would expect the control to "push away" from the boundary of G, and if this did not occur, then an enlargement of G might be called for.

2.3. The cost function. For $\alpha_i > 0$, $\beta_i > 0$, and $\beta > 0$, define the cost function

$$\begin{aligned} V(X(0), K, W) &= E\int_0^\infty e^{-\beta t}k(X(t))dt, \\ &+ E\int_0^\infty e^{-\beta t}[\alpha_1 dK^{12}(t) + \alpha_2 dK^{21}(t) + \beta_1 dU^1(t) + \beta_2 dU^2(t)], \qquad (2.3) \\ V(x) &= \inf V(x, K, W), \end{aligned}$$

where the infimum is over all admissible pairs. Assume the following.

Assumption 2.1. $b(\cdot)$, $\sigma(\cdot)$, and $k(\cdot)$ are continuous on G.

We write the cost as $V(X(0), K, W)$, since it is determined by the joint distribution of $X(0)$, $K(\cdot)$, and $W(\cdot)$. Since we do not require the control to be of the "feedback" form, we cannot write the cost as a function of $K(\cdot)$ only. In the "physical problem" leading to (2.1)–(2.3), the $U^i(\cdot)$ in (2.3) "penalizes" customers lost due to full buffers.

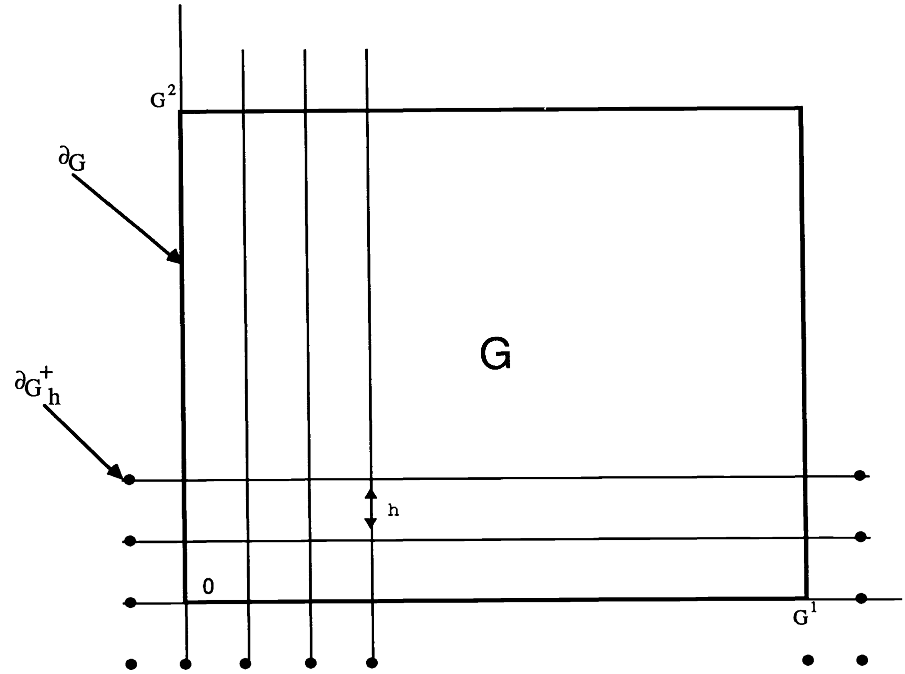


FIG. 1. *The numerical grid.*

3. A Markov chain approximation. The basic idea behind the numerical method is to find a discrete parameter Markov chain and associated control problem that is readily solvable and that approximates system (2.1)–(2.3) under the optimal control, in a sense to be described. The method is like that of [10], [11], but due to the presence of the singular control, some modifications of the technique of these references are required. Let h be an approximation parameter (h is a scalar here, but it could be vector valued; for example, in Fig. 1, the grid spacing could depend on the direction). Below, we will define a controlled Markov chain $\{\xi_n^h,\ n < \infty\}$, an admissible class of controls, and an associated control problem that provides an approximation to (2.1)–(2.3) in the following sense: An appropriate continuous parameter interpolation of the sequence of controlled chains $\{\xi_n^h,\ n < \infty\}$ converges weakly to a process satisfying (2.1) with an admissible pair $(K(\cdot), W(\cdot))$, as $h \to 0$. Let $V^h(x)$ denote the optimal value function for the control problem for the chain when $\xi_0^h = x$. Also, the continuous parameter interpolation of the *optimally controlled chain* converges weakly to the *optimally controlled* $X(\cdot)$ process and $V^h(x) \to V(x)$, for $x = X(0) \in G$. The numerical method thus consists in solving the optimal control problem for the chain. The particular chain that is used is chosen for convenience in solving the associated optimal control problem. The general approach has been used frequently. For the more classical stochastic control problems of [10], [11], there are standard and easily programmable methods for choosing the transition functions for the chain and for solving the control problem. Some modifications are required for the singular control problem and these will be discussed next.

To facilitate the development, we use a simple form for the state space, the G_h^+ defined below. Other forms will be discussed later, but the guiding ideas remain the same.

Define the h-grid G_h on G by $G_h = \{(kh, \ell h): 0 \le k \le G^1/h,\ 0 \le \ell \le G^2/h\}$, where we assume without loss of generality that the G^i are integral multiples of h. Define the "extended h-grid" $G_h^+ = \{(kh, \ell h): -1 \le k \le G^1/h{+}1,\ -1 \le \ell \le G^2/h{+}1\}$. Define $\partial G_h = G_h \cap \partial G$, the grid points on the boundary of G, and the "reflecting boundary" $\partial G_h^+ = G_h^+ - G_h$. See Fig. 1. The extended grid G_h^+ and ∂G_h^+ are not necessary but are introduced to allow us to deal with the reflection terms in a computationally efficient way. For states in the set G_h, the transition function for the Markov chain will be chosen such that the chain "behaves like" a controlled diffusion. When the state of the chain reaches ∂G_h^+ (i.e., has left G), it will be reflected back to G in a way that is consistent with the properties of the reflection terms $(I - P')Y - U$ in (2.1) (see Case 3 below). If we did not introduce ∂G_h^+, then the transition function would be more complex for states on ∂G. Define $\delta\xi_n^h = \xi_{n+1}^h - \xi_n^h$ and let E_n^h denote the expectation given all the data up to step n.

3.1. A heuristic discussion of the properties of the chain. Loosely speaking, the terms $J(\cdot)$, $Y(\cdot)$, and $U(\cdot)$ act either "instantaneously" or "impulsively." Intuitively, at the "instants" when $J(\cdot)$ changes there is no change in the $\int b\,dt + \int \sigma\,dW$ term. Since the behavior of the chain is supposed to "mimic" that of (2.1), we suppose that, for $\xi_n^h = x \in G_h$, we can choose to apply or not to apply a control. If we do not apply a control, then (loosely speaking) the increment $\delta\xi_n^h$ is to "behave like" an increment of $\int b\,dt + \int \sigma\,dW$ over a small time interval. We call this a "diffusion step." If we do apply a control at step n then, for consistency with the behavior of (2.1) at a "control instant," we will have

$$\delta\xi_n^h = (v_1 \delta K_n^{h,12} + v_2 \delta K_n^{h,21}) + (\text{small error}) \tag{3.1}$$

for some nonnegative control increments $\delta K_n^{h,ij}$ (only one of which will be nonzero). For programming simplicity and without loss of generality, we will restrict the $\delta K_n^{h,ij}$ to certain convenient values. See Case 2 below.

We can now write

$$\begin{aligned}\delta\xi_n^h = \delta\xi_n^h I_{\{\text{diffusion step at } n\}} &+ \delta\xi_n^h I_{\{\text{control step at } n\}} \\ &+ \delta\xi_n^h I_{\{\text{reflection step at } n\}}.\end{aligned} \tag{3.2}$$

The chain and the control will be chosen so that only one of the terms in (3.2) is positive. The transition function will be chosen so that the second term on the right will be (3.1). The third term on the right will have the form (see Case 3 below)

$$(I - P')\delta Y_n^h - \delta U_n^h + \text{"small error"} \tag{3.3}$$

for appropriate δY_n^h and δU_n^h, and the first term on the right will be formally similar to an increment in $\int b\,dt + \int \sigma\,dW$ over a small time interval.

The transition function $p^h(x, y \mid \text{control})$ for the chain is sometimes written as $p^h(x, y)$ if no control is used. We now define appropriate transition functions.

3.2. The transition functions.

Case 1. Let $\xi_n^h = x \in G_h$ and suppose that no control is exerted. Then (following the method in [10], [11]) the transition function is chosen such that the first two moments of $\delta\xi_n^h$ are "close" to those of $\int b\,dt + \int \sigma\,dW$ over a small time interval. In particular, define $a(x) = \sigma(x)\sigma'(x)$. Then for some $c_0 > 0$ and some "interpolation

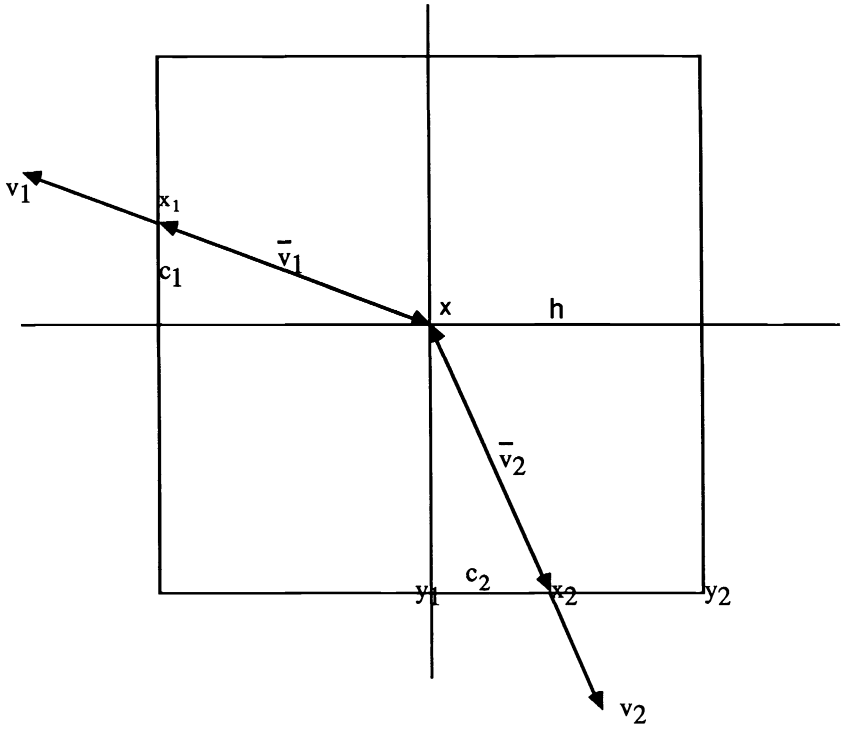


FIG. 2. *The control directions.*

interval" $\Delta t^h(x)$ satisfying $0 < c_0 h^2 \le \Delta t^h(x) \to 0$, we suppose that

$$
\begin{aligned}
&E_n^h \delta\xi_n^h = b(x)\Delta t^h(x) + O(h^\rho \Delta t^h(x)), \quad \rho > 0,\\
&\delta M_n^h = \delta\xi_n^h - E_n^h \delta\xi_n^h,\\
&E_n^h \delta M_n^h (\delta M_n^h)' = a(x)\Delta t^h(x) + O(h^\rho \Delta t^h(x)),\\
&|\xi_{n+1}^h - \xi_n^h| = O(h).
\end{aligned} \tag{3.4}
$$

The required p^h can be readily constructed. Some among the many possibilities are in [10], [11].

Case 2. *The control step.* Suppose that $x = \xi_n^h \in G_h$ and we decide to exert control, thus $\Delta t^h(x) = 0$. Define $\delta K_n^h = (\delta K_n^{h,12}, \delta K_n^{h,21})$ and $\delta J_n^h = v_1 \delta K_n^{h,12} + v_2 \delta K_n^{h,21}$. The transition probability $p^h(x, y \mid \text{control})$ will be chosen to satisfy (3.1). The "impulsive" control action is determined by a choice of the direction (either v_1 or v_2) together with the "impulse size" in that direction. For programming simplicity, we limit the choice of the nonnegative increments $\delta K_n^{h,ij}$ in the following way. Refer to Fig. 2 where $x = (x^1, x^2)$. Let x_i denote the intersection of the direction vectors v_i with the grid lines. Set $\bar{v}_i = x_i - x$. We restrict the δK so that the δJ take values either $\bar{v}_1$ or $\bar{v}_2$. Thus $\delta K_n^{h,ij}$ equals either 0 or h.

We restrict the values of $\delta K_n^{h,ij}$ since the method is easier to program if the choices are fewer and "local." Allowing arbitrary values for $\delta K_n^{h,ij}$ would yield the same limit results as in §5.

The points $x + \bar{v}_i = x_i$ are not usually in the grid G_h, unless the c_i equal unity. The actual transition function is chosen by a randomization so that the mean increment

$E_n^h \delta\xi_n^h$ is either $\bar{v}_1$ or $\bar{v}_2$, according to the choice of control action. In particular, if $\bar{v}_2$ is the chosen mean increment, then $\delta K_n^h = (0, h)$ and we use the transition function

$$p^h(x, y_1 \mid (0,h) = \delta K) = 1 - c_2 = 1 - p^h(x, y_2 \mid (0,h) = \delta K), \tag{3.5}$$

$$y_1 = (x^1, x^2 - h), \qquad y_2 = (x^1 + h, x^2 - h).$$

An analogous choice is used under $\bar{v}_1$. In the next section, we interpolate the $\{\xi_n^h\}$ into a continuous parameter process $\xi^h(\cdot)$. If at some n there is a control action, then the interpolation time is zero. In this sense, the original control problem is approximated by one in which the control acts impulsively.

Remark. The randomization is a perfectly natural numerical procedure. In fact, it is equivalent to a finite element approximation of $V(\cdot)$, i.e., a piecewise linear approximation. See the discussion in [11, §6] or the remark in Case 3 below.

By definition,

$$E_n^h \delta\xi_n^h = \delta J_n^h. \tag{3.6}$$

Let us define $\delta \tilde{J}_n^h = O(h)$ by

$$\delta\xi_n^h = \delta J_n^h + \delta \tilde{J}_n^h. \tag{3.7}$$

Then

$$E \sup_{n \le N} \left| \sum_{j=0}^{n-1} \delta \tilde{J}_j^h \right|^2 = NO(h^2) = O(h) E \sum_{j=0}^{N-1} |\delta J_j^h|. \tag{3.8}$$

In Theorem 5.3, (3.8) will be used to show that the $\{\delta \tilde{J}_n^h\}$ are asymptotically negligible, which implies that only the mean directions contribute to the limit, thus justifying the randomizations. The $\delta \tilde{J}_n^h$ are the "small errors" in (3.1). In analogy to (2.3), if $\delta K_n^{h,ij} \neq 0$ at step n, then the cost realized in this control step is $\alpha_i h = \alpha_i \delta K_n^{h,ij}$.

Case 3. *The reflection step.* For programming purposes, it seems simpler to separate the "diffusion" and "reflection" components of the chain approximation to (2.1) for $x \in \partial G$. This is the reason for the introduction of the separate "reflecting boundary" ∂G_h^+. We now discuss the reflection step for $x = \xi_n^h \in \partial G_h^+$. A useful method is described in [11, end of §9] and in [17] and we apply it to our problem here. The idea is similar to that of Case 2, namely, choose a mean reflection direction and state increment "consistent" with (2.1), and then "randomize" so that the actual new state is on the set of grid points ∂G_h. Let $x = (x^1, x^2) = \xi_n^h \in \partial G_h^+$. Refer to Fig. 3.

3.3. The transition function for the reflection step: the formal definition. We now fill in the details of the above heuristic discussion and obtain analogues of the U and Y terms needed for the convergence analysis in §5. There are three cases to be distinguished, depending on whether the components of $\xi_n^h = x = (x^1, x^2) \in \partial G_h^+$ are "too large," "too small," or a combination of these two: Case 3(i), all $x^i \ge 0$, some $x^i > G^i$ (the set of points between (a, b, c) in Fig. 3); Case 3(ii), all $x^i \le G^i$, some $x^i < 0$ (the set of points between (e, g, j)); Case 3(iii), some $x^i < 0$, some $x^j > G^j$ (the points d and k).

Case 3(i). Take $x = \xi_n^h \in \partial G_h^+$ to the nearest point on G_h, and define $\delta U_n^h = (\delta U_n^{h,1}, \delta U_n^{h,2}) = -\delta \xi_n^h$.

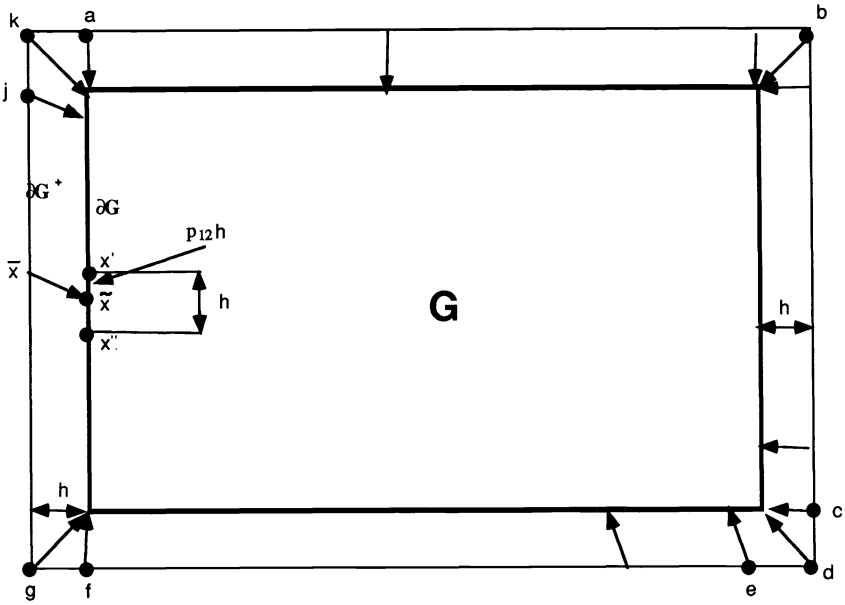


FIG. 3. *The reflection directions* ∂G.

Case 3(ii). To be consistent with the properties of the reflection term $(I - P')Y$ in (2.1), we first find a vector $\delta Y_n^h = (\delta Y_n^{h,1}, \delta Y_n^{h,2})$ with nonnegative components and the following properties: $\delta Y_n^{h,i} = 0$; if $x^i \geq 0$, the point $x + (I - P')\delta Y_n^h = \tilde{x}$ is on ∂G_h; if $x^i < 0$, then $\tilde{x}^i = 0$. This procedure yields a "consistent" reflection direction. However, $\tilde{x}$ will not generally be on ∂G_h. Refer to Fig. 3, and consider the point $\bar{x}$. Let x' and x'' denote the "neighboring" grid points to $\tilde{x}$, i.e., $\tilde{x} \in [x', x'']$. (In higher dimensions, we use a minimal set of points in ∂G_h such that $\tilde{x}$ is in its convex hull.) Define the transition function $p^h(\bar{x}, x') = |\tilde{x} - x''|/h = 1 - p^h(\bar{x}, x'')$. Thus, we randomize among x' and x'' keeping the mean value $\tilde{x}$. For the example in Fig. 3 where $x = \bar{x}$, we have $\delta Y_n^{h,2} = 0$, since $x^2 > 0$, and also

$$\tilde{x} - \bar{x} = (I - P') \begin{pmatrix} \delta Y_n^{h,1} \\ 0 \end{pmatrix} = \begin{pmatrix} h \\ 0 \end{pmatrix},$$

$$h = \delta Y_n^{h,1}, \qquad p^h(\bar{x}, x'') = p_{12}.$$

Case 3(iii). This is a combination of Cases 3(i) and 3(ii), since one component of x is "too big" and one is "too small." First, δY_n^h is defined, as in Case 3(ii), and then a δU_n^h is defined. To see the simple procedure, let $\xi_n^h = x =$ point d in Fig. 3, where $x^1 > G^1$ and $x^2 < 0$. Choose $\delta Y_n^{h,1} = 0$, $\delta Y_n^{h,2} \geq 0$ such that the second component of $\tilde{x} = x + (I - P')\delta Y_n^h$ is 0. The first component of $\tilde{x}$ will be greater than G^1 unless $p_{21} = 1$. Now, choose $\delta U_n^{h,1} = \tilde{x}^1 - G^1$, $\delta U_n^{h,2} = 0$. For reference in the dynamic programming equation (3.13) below, note that $\delta U_n^{h,1} = (1 - p_{21})h$ in this example. If $\xi_n^h =$ point k in the figure, then $\delta U_n^{h,1} = 0$ and $\delta U_n^{h,2} = (1 - p_{12})h$.

The method just described is indeed the "general method" for arbitrary state spaces. We calculate the intersection on ∂G of the line from $x \in \partial G_h^+$ in the reflection direction, and then randomize such that the mean point of intersection is preserved.

3.4. Representation of $\delta\xi_n^h$ in the reflection step. The following representation will be needed in §5.

In general, by the randomization used in Case 3, there is $\delta\tilde{Y}_n^h$ such that for $\xi_n^h \in \partial G_h^+$,

$$\begin{aligned} &\delta\xi_n^h = (I - P')\delta Y_n^h + \delta\tilde{Y}_n^h - \delta U_n^h, \\ &E_n^h \delta\tilde{Y}_n^h = 0, \qquad \delta\tilde{Y}_n^h = O(h), \end{aligned} \tag{3.9}$$

and the components of δY_n^h and $\delta\tilde{Y}_n^h$ (respectively, δU_n^h) can be nonzero only if $\xi_n^{h,i} < 0$ ($> G^i$, respectively). Define

$$U_n^h = \sum_{i=0}^{n-1} \delta U_i^h, \quad Y_n^h = \sum_{i=0}^{n-1} \delta Y_i^h, \quad \tilde{Y}_n^h = \sum_{i=0}^{n-1} \delta\tilde{Y}_i^h, \quad K_n^h = \sum_{i=0}^{n-1} \delta K_i^h.$$

We use K^h to denote the control policy. By the centering about conditional expectations, we have

$$\begin{aligned} E_n^h \left| \sum_{i=n}^{n+m-1} \delta\tilde{Y}_i^h \right|^2 &= O(h^2) E_n^h \quad (\#\text{ of reflection steps in } [n, n+m)) \\ &= O(h) E_n^h |Y_{n+m}^h - Y_n^h| \leq O(h^2) m. \end{aligned} \tag{3.10}$$

It will be shown in Theorem 5.3 that $E|Y_n^h|$ is bounded uniformly in h over the time interval of interest. Hence, formula (3.10) implies that the contributions of the $\{\delta\tilde{Y}_n^h\}$ terms vanish as $h \to 0$. Thus, the "randomization" in the reflection step has no effect in the limit.

3.5. The control problem for the chain. We say that a control policy K^h is *admissible* for the chain if it preserves the Markov property

$$P\{\xi_{n+1}^h = y \mid \xi_i^h, \ \delta K_i^h, \ i \leq n, \ \xi_n^h = x\} = p^h(x, y \mid \delta K_n^h).$$

Define $\Delta t_n^h = \Delta t^h(\xi_n^h)$ if n is a diffusion step, and set $\Delta t_n^h = 0$ otherwise. Define $t_n^h = \sum_{i=0}^{n-1} \Delta t_i^h$. With initial condition $\xi_0^h = x$ and control policy K^h, the cost function for the chain is defined to be

$$\begin{aligned} V^h(x, K^h) &= E \textstyle\sum_{n=0}^{\infty} e^{-\beta t_n^h} k(\xi_n^h) \Delta t_n^h \\ &+ E \textstyle\sum_{n=0}^{\infty} e^{-\beta t_n^h} [\alpha_1 \delta K_n^{h,12} + \ \alpha_2 \delta K_n^{h,21} + \beta_1 \delta U_n^{h,1} + \beta_2 \delta U_n^{h,2}]. \end{aligned} \tag{3.11}$$

This cost function is analogous to (2.3), in the sense that the sums in (3.11) "look like" integrals, due to the definition of the interpolation intervals. For $\xi_0^h = x$, define

$$V^h(x) = \inf_{K^h} V^h(x, K^h),$$

where the infimum is over all admissible policies.

3.5.1. The dynamic programming equation for the chain. The programming equation for (3.11) is the following. Write $p^h(x, y \mid \bar{v}_i)$ for the transition probability if the control is an increment in the direction v_i. For $x \in G_h$,

$$\begin{aligned} V^h(x) = \min \Big\{ & e^{-\beta \Delta t^h(x)} \textstyle\sum_y p^h(x, y) V^h(y) + k(x) \Delta t^h(x), \\ & \min_i \Big[\textstyle\sum_y p^h(x, y \mid \bar{v}_i) V^h(y) + \alpha_i h \Big] \Big\}. \end{aligned} \tag{3.12}$$

For $x \in \partial G_h^+$

$$V^h(x) = \textstyle\sum_y p^h(x,y)V^h(y) + \beta_1 h I_{\{x^1>G^1,x^2\geq 0\}} \tag{3.13}$$
$$+\ \beta_1 h(1-p_{21})I_{\{x^1>G^1,x^2<0\}} + \beta_2 h I_{\{x^2>G^2,x^1\geq 0\}}$$
$$+\ \beta_2 h(1-p_{12})I_{\{x^2>G^2,x^1<0\}}.$$

The reason for the $(1-p_{ij})$ terms is given in the discussion of Case 3(iii) above.

Particularly efficient methods for solving (3.12), (3.13) will be discussed in a forthcoming report. However, they can be solved by any of the usual iteration in policy or value space methods. In (3.12), there is a choice between no control and control, and a choice of the two alternatives for the control. There is no discount factor in the "control" term of (3.12), since $\Delta t_n^h = 0$ at any step at which there is a control action. A similar consideration holds for the reflection term (3.13).

3.5.2. A note on computation. In computing via the approximation in value space method, it has been observed that the rate of convergence of the iterates to $V^h(x)$ is much slower for $x \in \partial G_h^+$ than for $x \in G_h$. While we do not understand the reasons for this phenomenon, it should be kept in mind in setting the stopping criterion for the iterations.

3.5.3. Extensions. The method for the higher-dimensional problem $(r > 2)$ should be clear from the foregoing. The basic technique for dealing with the control and boundary terms is the same.

3.6. A formal dynamic programming (Bellman) equation for (2.1), (2.3). We now make some *purely formal* comments concerning the relationship between the numerical algorithm and a finite difference approximation to the Bellman equation for (2.1), (2.3). The comparison will provide a better intuitive feeling for the control step calculations. Write the *formal* Bellman equation, ignoring the boundary terms: Let $\mathcal{L}$ be the differential operator of the "diffusion" part of (2.1). Then (see [2] for a related problem)

$$0 = \min\left\{\mathcal{L}V(x) + k(x) - \beta V(x),\ \min_i (V_x'(x)v_i + \alpha_i)\right\}. \tag{3.14}$$

Note that the form of (3.14) implies a choice of "diffusion" or "control." A formal derivation of (3.14) can be obtained by writing $K^{ij}(\cdot)$ in the form $K^{ij}(t) = \int_0^t u^{ij}(s)ds$, with $0 \leq u^{ij}(s) \leq \infty$, and deriving the Bellman equation for the nonsingular problem. It will turn out that $u^{ij}(s)$ is either zero or infinity.

3.6.1. A finite difference interpretation of (3.12), (3.13). With appropriate choices of $p^h(x,y)$ for the diffusion step (particularly those based on the "finite difference model" in [10], [11]), the "diffusion part" of (3.11) can be rewritten in terms of a formal finite difference approximation to $\mathcal{L}V(x) + k(x)$ in (3.14). We now do the analogue for the control term in (3.14). Let e_i denote the unit vector in the ith coordinate direction and note that

$$1 - c_1 = p^h(x, x - e_1 h \mid \delta K = (h,0)),$$
$$1 - c_2 = p^h(x, x - e_2 h \mid \delta K = (0,h)).$$

Subtract $V^h(x)$ from both sides of (3.12). Then the inner minimum in (3.12), divided by h, equals $(i \neq j)$

$$\min_i \left[\frac{(V^h(x-e_ih) - V^h(x))}{h}(1-c_i) + \frac{(V^h(x-e_ih+e_jh) - V^h(x))}{h}c_i + \alpha_i\right]$$

$$\approx \min_i [V^h_{x_i}(x) - c_i V^h_{x_j}(x) + \alpha_i] = \min_i [V^h_x(x)' v_i + \alpha_i],$$

which is just the corresponding term in (3.14) with the superscript h dropped.

4. Interpolation and rescaling.

4.1. Preliminary results. The methods that have been used to show $V^h(x) \to V(x)$ for the more classical problems in [10], [11] need modification before they can be applied to the singular control problem. We first define some standard interpolations. Then we show why we might not have tightness and weak convergence in the Skorokhod topology and, finally, we define the rescaling that is used to obtain the desired limit results. The problem will also be put into a form that will be used in the convergence proofs of §5. Define $\xi^h(\cdot)$ by $\xi^h(t) = \xi^h_n$ on $[t^h_n, t^h_{n+1})$ for all n in which a diffusion step is used. Since $t^h_{n+1} = t^h_n$ if n is a control or reflection step, this defines $\xi^h(t)$ for all t. Define the process

$$K^{h,ij}(t) = \sum_{n:t^h_{n+1} \le t} \delta K^{h,ij}_n. \tag{4.1}$$

Define $B^h(\cdot)$, $M^h(\cdot)$, $Y^h(\cdot)$, $U^h(\cdot)$, $\tilde{Y}^h(\cdot)$, and $\tilde{J}^h(\cdot)$ by (4.1), but using (respectively) the sequences $\{b(\xi^h_n)\Delta t^h_n\}$, $\{\delta M^h_n\}$, $\{\delta Y^h_n\}$, $\{\delta U^h_n\}$, $\{\delta \tilde{Y}^h_n\}$, and $\{\delta \tilde{J}^h_n\}$. Define $J^h(t) = v_1 K^{h,12}(t) + v_2 K^{h,21}(t)$.

Now write

$$\begin{aligned} \xi^h_n &= x + \sum_{i=0}^{n-1} \delta\xi^h_i I_{\{\text{diffusion step at } i\}} \\ &\quad + \sum_{i=0}^{n-1} \delta\xi^h_i I_{\{\text{control step at } i\}} + \sum_{i=0}^{n-1} \delta\xi^h_i I_{\{\text{reflection step at } i\}}. \end{aligned}$$

Using representations (3.4), (3.7), (3.9), and the definitions of the above interpolations, the above formula yields (for $\xi^h(0) = x$)

$$\begin{aligned} \xi^h(t) &= x + B^h(t) + M^h(t) + J^h(t) + (I - P')Y^h(t) \\ &\quad - U^h(t) + \tilde{Y}^h(t) + \tilde{J}^h(t) + \textstyle\sum_{i:t^h_i \le t} O(h^\rho \Delta t^h_i), \end{aligned} \tag{4.2}$$

where

$$B^h(t) = \int_0^t B(\xi^h(s))ds + O(h).$$

Also, $\{\sum_{i=0}^{n-1} \delta M^h_i\}$ is a martingale and

$$E^h_t[M^h(t+s) - M^h(t)][M^h(t+s) - M^h(t)]' = E^h_t \int_t^{t+s} a(\xi^h(s))ds + O(h^\rho),$$

where E^h_t denotes the expectation given the data up to interpolated time t.

Approximating $e^{-\beta t^h_n}$ by $e^{-\beta t}$ on $[t^h_n, t^h_{n+1})$, for an admissible policy K^h, we have the representation for the cost

(4.3)

$$\begin{aligned} V^h(x, K^h) &= E \textstyle\int_0^\infty e^{-\beta s} k(\xi^h(s))ds \\ &\quad + E \textstyle\int_0^\infty e^{-\beta s}[\alpha_1 dK^{h,12}(s) + \alpha_2 dK^{h,21}(s) + \beta_1 dU^{h,1}(s) + \beta_2 dU^{h,2}(s)]. \end{aligned}$$

The similarity of (4.2), (4.3) with (2.1), (2.3) suggests that our continuous parameter interpolations are appropriate.

4.2. A difficulty with the weak convergence. To prove that $V^h(x) \to V(x)$, we would like to use the weak convergence technique of [10] or [11]. By that method, we first prove tightness of the set of processes in (4.2). We then extract a weakly convergent subsequence and show that the limit satisfies (2.1) for some admissible pair $(K(\cdot), W(\cdot))$, where $K(\cdot) = \lim_h K^h(\cdot)$, and that either $V^h(x) \to V(x, K, W)$ or $\underline{\lim}_h V^h(x) \geq V(x, K, W)$. This implies that $\underline{\lim}_h V^h(x) \geq V(x)$. The reverse inequality $\overline{\lim}_h V^h(x) \leq V(x)$ is then proved by a separate argument.

This method does not work here, since we do not know whether $\{K^h(\cdot)\}$ is tight. Owing to the time scaling, it is possible to alternate diffusion and control steps in such a way that the "limit" has a "jump," and there is no convergence in the Skorokhod topology.

There are several ways to handle this "singular" problem. The pseudopath topology that was used in [18] for a weak convergence study of a singular control problem could be used. We prefer to use a time change argument of the type used in [7] since it is more direct. The idea is to rescale time so that tightness is assured, take weak limits, and then use an inverse transformation to obtain the desired result. The method is quite useful for the study of limits of sequences of singularly controlled processes. In preparation for this work in the next section, we introduce the following rescaling.

4.3. The rescaled processes. Define $\Delta\hat{t}_n^h$ by $\Delta\hat{t}_n^h = \Delta t_n^h$ if step n is a "diffusion" step, $\Delta\hat{t}_n^h = 0$ if we reflect on step n, and $\Delta\hat{t}_n^h = h$ if we control at step n. Set $\hat{t}_n^h = \sum_{i=0}^{n-1} \Delta\hat{t}_i^h$. Define $\hat{T}^h(\cdot)$ by $\hat{T}^h(0^-) = 0$, and $\hat{T}^h(t) = \sum_{i=0}^{n} \Delta t_i^h = t_{n+1}^h$ on the interval $[\hat{t}_n^h, \hat{t}_{n+1}^h]$. Thus $\hat{T}^h(\cdot)$ does not increase at these t at which a control step occurs.

Define the rescaled and interpolated processes $\hat{\xi}^h(\cdot)$ and $\hat{M}^h(\cdot)$ by $\hat{\xi}^h(t) = \xi_n^h$ on $[\hat{t}_n^h, \hat{t}_{n+1}^h)$ and $\hat{M}^h(t) = \sum_{i=0}^{n-1} \delta M_i^h$ on $[\hat{t}_n^h, \hat{t}_{n+1}^h)$, and define $\hat{K}^h(\cdot)$, $\hat{J}^h(\cdot)$, $\hat{Y}^h(\cdot), \cdots$, analogously to $\hat{M}^h(\cdot)$. We can write

$$\begin{aligned}\hat{\xi}^h(t) = x + \hat{B}^h(t) + \hat{M}^h(t) + \hat{J}^h(t) + (I - P')\hat{Y}^h(t) - \hat{U}^h(t) \\ + \hat{\hat{Y}}^h(t) + \hat{\tilde{J}}^h(t) + (\text{negligible error}),\end{aligned} \tag{4.4}$$

$$\begin{aligned}V^h(x, K^h) = E \textstyle\int_0^\infty e^{-\beta\hat{T}^h(s^-)} k(\hat{\xi}^h(s)) d\hat{T}^h(s) \\ + E \textstyle\int_0^\infty e^{-\beta\hat{T}^h(s^-)} [\alpha_1 d\hat{K}^{h,12}(s) \\ + \alpha_2 d\hat{K}^{h,21}(s) + \beta_1 d\hat{U}^{h,1}(s) + \beta_2 d\hat{U}^{h,2}(s)].\end{aligned} \tag{4.5}$$

Comment. The rescaling "stretches out" the control and state processes, so that they are smoother and tightness can be proved. In fact, the $\hat{K}^{h,ij}(\cdot)$ are Lipschitz continuous with constant unity for all ω. In the next section, the weak convergence analysis is done for the rescaled process. Then, via an inverse time transformation of the limit process, $V^h(x) \to V(x)$ is obtained. The time rescaling arguments parallel those in [7] as closely as possible, but the notation and many details are different, since in [7], we worked with a "physical" heavy traffic process with a different structure.

5. The convergence theorem. Theorem 5.3 proves the weak convergence of the rescaled processes and gives a representation of the limit in terms of a rescaled form of (2.1). It also shows that by a natural inverse scaling, we recover a process of the form (2.1). Theorem 5.4 proves that $V^h(x)$ converges to the cost for some process of the type (2.1). Theorem 5.6 completes the proof that $V^h(x) \to V(x)$.

We next quote two results that will be needed in the proof of Theorems 5.3 and 5.5.

THEOREM 5.1 (The reflection mapping). *Let P be a degenerate Markov transition matrix, with spectral radius less than unity. Let $z(\cdot) \in D^k[0,\infty)$ and consider the equation*

$$x(t) = z(t) + (I - P')y(t) - u(t), \quad x^i(t) \in [0, G^i] \text{ for all } t. \tag{5.1}$$

There is a unique continuous function (in the topology of uniform convergence on bounded time intervals) $F(\cdot)$ such that $(y(\cdot), u(\cdot)) = F(z(\cdot))$ has the following properties: $F(\cdot)$ maps $C^k[0,\infty)$ into $C^k[0,\infty)$ and $D^k[0,\infty)$ into $D^k[0,\infty)$; the $y^i(\cdot)$ (respectively, $u^i(\cdot)$) are nondecreasing and nonnegative and can increase only when $x^i(t) = 0$ (respectively, $x^i(t) = G^i$), and $y(0) = u(0) = 0$. Also, (5.1) *holds.*

The proof is in [7, §8] and is an extension of a similar theorem in [16].

THEOREM 5.2. *Assume that Assumption* 2.1 *holds true. Let $V^h(x,0)$ denote the cost with no control used. Then $\sup_{x,h} V^h(x,0) < \infty$. Let $\{K^{h,ij}(n+1) - K^{h,ij}(n), h, n\}$ be uniformly integrable. Then so is $\{U^h(n+1) - U^h(n),\ h, n\}$.*

The proof is the same as that of [7, Thms. 7, 9]; just replace the $(M^\epsilon, B^\epsilon, B_i)$ there by our (M^h, B^h, G^i).

Define the process $\hat{H}^h(\cdot)$ by

$$\hat{H}^h(\cdot) = (\hat{\xi}^h(\cdot), \hat{M}^h(\cdot), \hat{B}^h(\cdot), \hat{K}^h(\cdot), \hat{\tilde{J}}^h(\cdot), \hat{Y}^h(\cdot), \hat{\tilde{Y}}^h(\cdot), \hat{U}^h(\cdot), \hat{T}^h(\cdot)).$$

The statement of Theorem 5.3 is a little long, but it seems preferable to have the results in one place.

THEOREM 5.3. *Assume that Assumption* 2.1 *holds true, and let $K^h(\cdot)$ be admissible control policies. Then $\{\hat{H}^h(\cdot)\}$ is tight and the limit of any weakly convergent subsequence is continuous with probability one (w.p.*1.*) Let $\hat{H}(\cdot)$ denote the limit of such a subsequence. Then $\hat{\tilde{Y}}(\cdot)$ and $\hat{\tilde{J}}(\cdot)$ are identically zero w.p.*1 *and ($z(0) = x$ here)*

$$\hat{\xi}(t) = x + \hat{M}(t) + \hat{B}(t) + \hat{J}(t) + (I - P')\hat{Y}(t) - \hat{U}(t), \tag{5.2}$$

where $\hat{J}(t) = v_1\hat{K}^{12}(t) + v_2\hat{K}^{21}(t)$, $\hat{B}(t) = \int_0^t b(\hat{\xi}(s))d\hat{T}(s)$, and $\hat{M}(\cdot)$ is a $\mathcal{B}(\hat{H}(s), s \leq t)$-martingale with quadratic variation process $\int_0^t a(\hat{\xi}(s))d\hat{T}(s)$.

*Assume, further, that $\sup_h E|K^h(t)| < \infty$ for each t. Then, $\hat{T}(t) \to \infty$ w.p.*1. *For $t < \infty$, define the inverse $T(\cdot)$ by $T(t) = \inf\{s: \hat{T}(s) > t\}$. Then $T(t) \to \infty$ w.p.*1 *as $t \to \infty$. For any process $\hat{\phi}(\cdot)$, define the rescaled process $\phi(\cdot)$ by $\phi(t) = \hat{\phi}(T(t))$. Then*

$$\xi(t) = x + M(t) + B(t) + J(t) + (I - P')Y(t) - U(t), \tag{5.3}$$

where $J(t) = v_1K^{12}(t) + v_2K^{21}(t)$ and $B(t) = \int_0^t b(\xi(s))ds$. Also, $M(\cdot)$ is a $\mathcal{B}_t = \mathcal{B}(H(s), s \leq t)$-martingale with quadratic variation $\int_0^t a(\xi(s))ds$. Finally, there is a $\mathcal{B}_t$-Wiener process $W(\cdot)$ such that $M(t) = \int_0^t \sigma(\xi(s))dW(s)$ and $(K(\cdot) - K(0), W(\cdot))$ is an admissible pair.

Remark. By the definition of $T(\cdot)$ and the continuity of $\hat{H}(\cdot)$, the components of $H(\cdot)$ have paths in $D[0,\infty)$. However, $T(\cdot)$, $\xi(\cdot)$, $K(\cdot)$, $U(\cdot)$, or $Y(\cdot)$ might be discontinuous due to an "accumulation" of control actions at some time point. Consider one case, where $\hat{T}(t) = 0$ on an interval $[0,\delta)$, $\delta > 0$. This can happen if control and diffusion steps alternate for the chain, with a reflection step taken where necessary. For this case, $K(0) \neq 0$. Consider another case, where there is an "accumulation" of control actions "pushing out of G" when $\xi_n^h \in \partial G_h$. Then $Y(\cdot)$ or $U(\cdot)$ would be discontinuous. If such "accumulations" occurred at $t = 0$, then there would be an impulsive change in the state at $t = 0$ of magnitude $K(0) + (I - P')Y(0) - U(0)$.

Proof. Part 1. *Tightness.* Let τ denote a finite stopping time. By the boundedness of $b(\cdot)$ and the martingale property of $\{M_n^h,\ n < \infty\}$, we have (the $O(\cdot)$ terms do not depend on τ)

$$E_\tau^h |B^h(\tau + s) - B^h(\tau)| \leq O(s) + O(h),$$
$$E_\tau^h |M^h(\tau + s) - M^h(\tau)|^2 \leq O(s) + O(h).$$

By [15, Thm. 2.7b], a sufficient condition for tightness of a sequence of processes $\{\phi_n(\cdot)\}$ with paths in $D[0, \infty)$ is that for each $T < \infty$

$$\lim_{s \to 0} \overline{\lim_{n}} \sup_{\tau \leq T} E|\phi_n(\tau + s) - \phi_n(\tau)| = 0, \tag{5.4}$$

where the supremum is over all stopping times less than T. Thus $\{B^h(\cdot),\ M^h(\cdot)\}$ is tight. Hence so is $\{\hat{B}^h(\cdot), \hat{M}^h(\cdot)\}$ due to the "stretching out" of the timescale. By the definition of the "stretched out" timescale, $|\hat{K}^h(t + s) - \hat{K}^h(t)| \leq |s| + O(h)$. Thus $\{\hat{K}(\cdot)\}$ is tight. Consequently, by (3.8), $\{\hat{\tilde{J}}^h(\cdot)\}$ converges weakly to the zero process.

Note that a reflection step can occur only after a diffusion or control step. Then (3.10) and the fact that $\Delta \hat{t}_n^h \geq c_0 h^2$ (see (3.4) above) for the control or diffusion step imply that for $s > 0$ and any bounded stopping time τ

$$\begin{aligned} E_\tau^h |\hat{\tilde{Y}}^h(\tau + s) - \hat{\tilde{Y}}^h(\tau)|^2 &= O(h^2) E_\tau^h \text{ (number of reflection steps of } \hat{\xi}^h(\cdot) \text{ in} \\ &\qquad \text{interpolated interval } [\tau, \tau + s)) \\ &= O(h^2) E_\tau^h \text{ (number of diffusion or control steps of } \hat{\xi}^h(\cdot) \\ &\qquad \text{in interpolated interval } [\tau, \tau + s)) \\ &= O(h^2)(s/h^2 + 1) = O(s) + O(h^2), \end{aligned}$$

which together with (5.4) yields the tightness of $\{\hat{\tilde{Y}}^h(\cdot)\}$. All the limits of weakly convergent subsequences of these processes are continuous w.p.1, since the jumps are all of order $O(h)$. We also have

$$E_\tau^h |\hat{\tilde{Y}}^h(\tau + s) - \hat{\tilde{Y}}^h(\tau)|^2 = O(h) E_\tau^h |\hat{Y}^h(\tau + s) - \hat{Y}^h(\tau)|.$$

Write (4.4) as

$$\hat{\xi}^h(t) = \hat{Z}^h(t) + (I - P')\hat{Y}^h(t) - \hat{U}^h(t),$$

where $\hat{Z}^h(\cdot)$ is defined in the obvious way from (4.4). We have proved that $\{\hat{Z}^h(\cdot)\}$ is tight and all the limits are continuous processes. By Theorem 5.1, there is a unique continuous function $F(\cdot)$ (not depending on h) that is continuous and is such that $(\hat{Y}^h(\cdot), \hat{U}^h(\cdot)) = F(\hat{Z}^h(\cdot))$ is the unique process for which $\hat{\xi}^h(t) \in G$ and where the components of $(\hat{Y}^h(\cdot),\ \hat{U}^h(\cdot))$ can increase only when $\hat{\xi}^h(\cdot)$ is on the appropriate part of the boundary. Now by the continuity of $F(\cdot)$, $\{\hat{Y}^h(\cdot), \hat{U}^h(\cdot)\}$ is tight and all weak limits are continuous. This and the above estimate on $\hat{\hat{Y}}^h(\cdot)$ imply that $\hat{\hat{Y}}^h(\cdot)$ converges weakly to the zero process. (Note that we defined $\hat{Y}^h(\cdot)$ and $\hat{U}^h(\cdot)$ a priori. The representation of Theorem 5.1 tells us that they can also be obtained as a function of $\hat{Z}^h(\cdot)$. This latter representation is useful to prove the tightness and continuity properties of $\hat{U}^h(\cdot)$ and $\hat{Y}^h(\cdot)$, as in [16], [17], [7].)

Part 2. *Weak convergence.* Abusing terminology, let h index a weakly convergent subsequence, with limit denoted by $\hat{H}(\cdot)$. The paths of $\hat{H}(\cdot)$ are continuous w.p.1, and $\hat{\hat{J}}(t) = \hat{\hat{Y}}(t) \equiv 0$. The limit satisfies (5.2). Since

$$\hat{B}^h(t) = \int_0^t b(\hat{\xi}(s))d\hat{T}^h(s) + O(h), \tag{5.5}$$

we have the representation $\hat{B}(t) = \int_0^t b(\hat{\xi}(s))d\hat{T}(s)$.

Recall that $\{M_n^h\}$ is a martingale and that

$$E_n^h[M_{n+m}^h - M_n^h][M_{n+m}^h - M_n^h]' = E_n^h \sum_{i=n}^{n+m-1} [a(\xi_i^h)\Delta t_i^h + O(h^\rho)\Delta t_i^h].$$

Let $\hat{E}_t^h$ denote the expectation conditioned on $\hat{\mathcal{B}}_t^h \equiv \mathcal{B}(\hat{H}^h(s),\ s \le t)$. For any bounded $\hat{\mathcal{B}}_t^h$-stopping time τ, we can write

$$\hat{E}_\tau^h[\hat{M}^h(\tau + s) - \hat{M}^h(\tau)] = 0,$$

$$\begin{aligned}\hat{E}_\tau^h&[\hat{M}^h(\tau + s) - \hat{M}^h(\tau)][\hat{M}^h(\tau + s) - \hat{M}^h(\tau)]' \\ &= \hat{E}_\tau^h \int_\tau^{\tau+s} a(\hat{\xi}^h(s))d\hat{T}^h(s) + O(h).\end{aligned} \tag{5.6}$$

Let q be an arbitrary integer, let $t_i \le t < s$, $i \le q$, and let $h(\cdot)$ be a bounded and continuous real-valued function of its arguments. Then (5.6) implies that

$$Eh(\hat{H}^h(t_i), i \le q)[\hat{M}^h(t + s) - \hat{M}^h(t)] = 0,$$

$$\begin{aligned}Eh(\hat{H}^h(t_i), i \le q)\Big(&[\hat{M}^h(t + s) - \hat{M}^h(t)][\hat{M}^h(t + s) - \hat{M}^h(t)]' \\ &- \int_t^{t+s} a(\hat{\xi}^h(s))d\hat{T}^h(s)\Big) = O(h^\rho).\end{aligned} \tag{5.7}$$

By taking limits in (5.7), we see that (5.7) holds with the h superscript dropped. The arbitrariness of $h(\cdot)$, q, t_i, t, and s imply that $\hat{M}(\cdot)$ is a $\hat{\mathcal{B}}_t$-martingale with the quadratic variation given in the theorem statement.

Part 3. *The construction of* $\xi(\cdot)$. Since $\sup_h E|K^h(t)| < \infty$ for each t by hypothesis, we have $\hat{T} \to \infty$ w.p.1 as $t \to \infty$. This is because, *loosely speaking,* $\hat{T}^h(t + |K^h(t)|) \approx t$. Thus the inverse function $T(s)$ defined in the theorem statement exists for each s. Note that for $u \ge 0$, we have $\{T(s) \le u\} = \{\hat{T}(u) \ge s\} \in \hat{\mathcal{B}}_u$. Thus

$T(s)$ is a $\hat{\mathcal{B}}_t$-stopping time for each s. Consequently, $M(\cdot)$ is a $\mathcal{B}_t \equiv \mathcal{B}(H(s),\ s \le t)$-martingale. The quadratic variation of $M(\cdot)$ is easily calculated to be $\int_0^t a(\xi(s))ds$. (This fact follows simply by rescaling the limit of (5.7).) Hence, there is a $\mathcal{B}_t$-Wiener process $W(\cdot)$ such that $M(t) = \int_0^t \sigma(\xi(s))dW(s)$. Clearly, $(K(\cdot), W(\cdot))$ is an admissible pair, since $\mathcal{B}_t$ measures $\{K(s),\ s \le t\}$ for each t. If $a(\cdot)$ is degenerate at some point, then we might have to augment the probability space by adding an "independent" Wiener process. It is also easily shown that $B(t) = \int_0^t b(\xi(s))ds$. Thus representation (5.3) follows from (5.2). □

The next theorem characterizes the limit of the costs $V^h(x, K^h)$ and gives "one half" of the desired convergence theorem for $\{V^h(x)\}$.

THEOREM 5.4. *Assume that Assumption* 2.1 *holds true and use the notation of Theorem* 5.3. *If* $\{K^h(n+1) - K^h(n),\ h, n\}$ *is uniformly integrable, then for the weakly convergent subsequence in Theorem* 5.3 (*indexed by* h),

$$V^h(x, K^h) \to V(x, K, W).$$

If the uniform integrability hypothesis is not satisfied, then

$$\underline{\lim}_h V^h(x, K^h) \ge V(x, K, W) \ge V(x).$$

Proof. By the weak convergence and the continuity of the limit $\hat{H}(\cdot)$,

$$V^h(x) = E\sum_{n=0}^{\infty} e^{-\beta t_n^h} k(\xi_n^h)\Delta t_n^h = E\int_0^{\infty} e^{-\hat{T}^h(s^-)} k(\hat{\xi}^h(s))d\hat{T}^h(s), \tag{5.8}$$

$$E\int_0^{\infty} e^{-\hat{T}^h(s^-)} k(\hat{\xi}^h(s))d\hat{T}^h(s) \to E\int_0^{\infty} e^{-\beta\hat{T}(s)} k(\hat{\xi}(s))d\hat{T}(s).$$

On inversion of the timescale, the right side of (5.8) can be written as $E\int_0^\infty e^{-\beta t} k(\xi(t))dt$. By the uniform integrability of $\{K^h(n+1) - K^h(n),\ h, n\}$ and Theorem 5.2, $\{U^h(n+1) - U^h(n),\ h, n\}$ is uniformly integrable. The right-hand sum in (3.11) can be written as

$$E\int_0^{\infty} e^{-\beta\hat{T}^h(s^-)}[\alpha_1 d\hat{K}^{h,12}(s) + \alpha_2 d\hat{K}^{h,21}(s) + \beta_1 d\hat{U}^{h,1}(s) + \beta_2 d\hat{U}^{h,2}(s)]. \tag{5.9}$$

By the weak convergence, continuity of the limit process, and uniform integrability, (5.9) converges to

$$E\int_0^{\infty} e^{-\beta\hat{T}(s)}[\alpha_1 d\hat{K}^{12}(s) + \alpha_2 d\hat{K}^{21}(s) + \beta_1 d\hat{U}^{1}(s) + \beta_2 d\hat{U}^{2}(s)].$$

On inversion of the timescale, the above expression can be written as

$$E\int_0^{\infty} e^{-\beta t}[\alpha_1 dK^{12}(t) + \alpha_2 dK^{21}(t) + \beta_1 dU^{1}(t) + \beta_2 dU^{2}(t)]. \tag{5.10}$$

The right-hand sides of (5.8) and (5.10) add up to $V(x, K, W)$. In the absence of the uniform integrability, we use Fatous' lemma to get $\underline{\lim}_h V^h(x, K^h) \ge V(x, K, W)$. □

THEOREM 5.5. *Assume that Assumption* 2.1 *holds true. Then*

$$\varliminf_{h} V^h(x) \geq V(x). \tag{5.11}$$

Proof. Let K_0^h denote an optimal admissible policy for $\{\xi_n^h, n < \infty\}$. By Theorem 5.2 and the discounted nature of the cost function, we have the following: For any $\delta > 0$, there is a $T > 0$ (independent of h) such that for each h there is a δ-optimal policy $K^h(\cdot)$ for which no control is exercised after T, i.e., for n such that $t_n^h \geq T$. Since $\sup_h V^h(x) < \infty$ by Theorem 5.2, we have $\sup_h |K^h(t)| < \infty$ for each t. Then the theorem follows from Theorem 5.4 and the fact that for any weakly convergent subsequence of $\{\hat{H}^h(\cdot)\}$ with limit $\hat{H}(\cdot)$ and with the rescaled process defined by $H(\cdot) = \hat{H}(T(\cdot))$, we have $V(x, K, W) \geq V(x)$, where $(K(\cdot), W(\cdot))$ is the admissible pair in $H(\cdot)$. □

To complete the proof that $V^h(x) \to V(x)$, we need to show that $\overline{\lim}_h V^h(x) \leq V(x)$. To do this, we need to exploit the optimality of $V^h(x)$. This is done in the next theorem.

DEFINITION. The solution to (2.1) is said to be *unique in the weak sense* if the probability law of $(K(\cdot), W(\cdot))$ determines that of $(X(\cdot), K(\cdot), W(\cdot))$. We will need the following assumption.

Assumption 5.1. For each $\delta > 0$, there is some δ-optimal control for (2.1), (2.3) for which the solution to (2.1) is unique in the weak sense. The uncontrolled system (2.1) (i.e., with $K(t) \equiv 0$) has a weak sense unique solution for each initial condition.

THEOREM 5.6. *Assume that Assumptions* 2.1 *and* 5.1 *hold true. Then* $V^h(x) \to V(x)$.

Proof. To exploit the fact that $V^h(x)$ is an optimal cost for the chain $\{\xi_n^h, n < \infty\}$, we choose a particular δ-optimal control for (2.1), (2.3), such that the approximation can be applied to the chain and the associated cost compared with $V^h(x)$. The theorem will follow from a comparison of the costs associated with appropriate weakly convergent subsequences. The chosen δ-optimal control is not a "practical" control, but it does give the desired inequalities. The proof is an adaptation of a method used for a "heavy traffic" problem in [7]. The proof of the existence of a δ-optimal control with the following properties is in [7, §6]. There are analogous proofs for other singular control problems. Let $\delta > 0$. Then under Assumptions 2.1 and 5.1 there is a δ-optimal admissible (with respect to the Wiener process $W(\cdot)$) control $K(\cdot)$ for (2.1), (2.2) with the following properties: (i) There are $T_\delta < \infty$, $\Delta > 0$, $\gamma > 0$, and $\rho > 0$ such that the $K^{ij}(\cdot)$ are constant on the intervals $[n\Delta, n\Delta + \Delta)$, only one of its components can jump at a time, and the jumps take values in the discrete set $k\rho$, $k = 1, 2, \cdots$; also, $K(\cdot)$ is bounded and is constant on $[T_\delta, \infty)$; (ii) the values are determined by the conditional probability laws (which defines the functions $q_{nki}(\cdot)$)

$$\begin{aligned} &P\{dK^{ij}(n\Delta) = k\rho \mid K(m\Delta), m < n, W(s), s \leq n\Delta\} \\ &\quad = P\{dK^{ij}(n\Delta) = k\rho \mid K(m\Delta), m < n, W(p\gamma), p\gamma \leq n\Delta\} \\ &\quad \equiv q_{nki}(K(m\Delta), m < n, W(p\gamma), p\gamma \leq n\Delta), \end{aligned} \tag{5.12}$$

and the $q_{nki}(\cdot)$ can be supposed to be continuous in the W-variables for each value of the control variables. Equation (5.12) says essentially that the conditional distribution of $dK(n\Delta)$, given the "past," equals the conditional distribution given only certain sampled values of the "past," and that it is a continuous function of these values.

We next adapt this control policy (5.12) to the chain $\{\xi_n^h\}$. To do this with minimal effort, assume that $\sigma^{-1}(x)$ exists for $x \in G$. The general case is handled by an "approximation" procedure. If step n is a "diffusion" step, then define

$$\delta W_n^h = \sigma^{-1}(\xi_n^h)[\delta \xi_n^h - E_n^h \delta \xi_n^h].$$

Define $W_n^h = \sum_{i=0}^{n-1} \delta W_i^h$ and $W^h(\cdot)$ by $W^h(t) = W_n^h$ on $[t_n^h, t_{n+1}^h)$. Note that

$$E_n^h[W_{n+m}^h - W_n^h] = 0,$$

$$E_n^h[W_{n+m}^h - W_n^h][W_{n+m}^h - W_n^h]' = E_n^h \sum_{i=n}^{n+m-1} \Delta t_i^h(1 + O(h^\rho))I,$$

where I is the identity matrix.

By a proof analogous to the weak convergence part of Theorem 5.3, it is easily shown that $\{W^h(\cdot)\}$ converges weakly to a standard vector-valued Wiener process $W(\cdot)$; in fact, this Wiener process can be used in the representation of the $M(\cdot)$ given at the end of Theorem 5.3.

The δ-optimal control $K(\cdot)$ defined by (5.12) is "impulsive," and we will adapt it for use on the chain. As preparation for the proof, we first note the following. Suppose that we wish to apply a control of "impulsive" magnitude δK^{ij} to the chain at some interpolated time t_0. Define $n_h = \min\{k: t_k^h \geq t_0\}$. Then starting at step n_h, apply $[\delta K^{ij}/h]$ successive control steps, each of the randomized type of Case 2 in §3 (with reflection steps intervening if a control step takes the chain out of G_h). As shown in Theorem 5.3, the effects of the randomization disappear as $h \to 0$. Let $K_0^h(\cdot)$ denote the continuous parameter interpolation (analogous to (4.1)) of the control just defined. We note further that $\{K_0^h(\cdot)\}$ is tight in the Skorohod topology, and the weak limit is just a step function of jump δK^{ij} at time t_0. When using such controls, there is no need to rescale time in the convergence proof, as was done in Theorem 5.3. The tightness holds because the values of the times t_n^h do not increase during the sequence of successive control steps just described, since $\Delta t_k^h = 0$ if k is a control or reflection step. Hence the interpolation $K^{h,ij}(\cdot)$ is just a step function with jump $h[\delta K^{ij}/h]$ at time $t_{n_h}^h$. Analogous arguments can be used if there are a finite number of such jumps at discrete times $\{t_k\}$.

With the observations in the last paragraph, we are ready to define the "adapted" form of $K(\cdot)$ for use on $\{\xi_n^h,\ n < \infty\}$. Let $K^h(\cdot)$ denote the interpolated form of the "adaptation." We will define $K^h(\cdot)$ such that it has the same number of impulsive changes as does $K(\cdot)$ (at most T_δ/Δ), each being uniformly bounded. Each of the impulses is to be realized for the chain via the method used in the example above. The impulses are to occur as soon after the "interpolated times" $k\Delta$ as possible. Let δK_n^h denote the value of the impulse that we would like to apply to the chain at interpolated time $n\Delta$. Define $\bar{t}_n^h = \min\{t_k^h: t_k^h \geq n\Delta\}$. Then the δK_n^h are chosen by the conditional probability rule

$$\begin{aligned} P\{\delta K_n^{h,ij} &= \rho k \mid \xi^h(s), s \leq \bar{t}_n^h, \delta K_m^h, m < n\} \\ &= q_{nki}(\delta K_m^h, m < n, W^h(p\gamma), p\gamma \leq n\Delta). \end{aligned}$$

The sequence $\{K^h(\cdot), W^h(\cdot)\}$ is tight. By construction, the weak limit has the distribution of the $(K(\cdot), W(\cdot))$ of (5.12). Thus we can denote the limit by $(K(\cdot), W(\cdot))$ without confusion.

By a weak convergence argument analogous to that of Theorem 5.3, but *without the time rescaling,* we get $\{\xi^h(\cdot),\ K^h(\cdot),\ W^h(\cdot),\ Y^h(\cdot),\ U^h(\cdot)\}$ converges weakly to a set $(\xi(\cdot),\ K(\cdot),\ W(\cdot),\ Y(\cdot),\ U(\cdot))$, solving (2.1).

By the weak sense uniqueness to the uncontrolled form of (2.1) for each initial condition, and the impulsive nature of the control, the solution to (2.1) under the "impulsive" δ-optimal control $K(\cdot)$ used here is also unique in the weak sense. Thus $\xi(\cdot)$ is the unique solution to (2.1) driven by the chosen δ-optimal control. Since $\{K^h(\cdot)\}$ is bounded, Theorem 5.4 implies that $V^h(x, K^h) \to V(x, K, W)$. By the optimality of $V^h(x)$ and the δ-optimality of $K(\cdot)$, $V(x, K, W) \le V(x) + \delta$ and $V^h(x, K^h) \ge V^h(x)$. Thus $\overline{\lim}_h V^h(x) \le V(x)$. This together with Theorem 5.5 yields the theorem. □

6. A minimum-fuel-type problem. In the previous sections, only two directions (v_1 and v_2) were allowed for the control actions (in addition to the possibility of no control) at each step. This limitation was due only to the structure of the original class of physical problems. To illustrate other possibilities, in this section we discuss a problem where the number of possible control directions is infinite. Again, a special case will be used to make the main point. The chosen case has an interesting additional feature due to the possibility of a rapidly varying control. Again, the problem is canonical in that it is representative of a large class.

We consider a form of a problem dealt with by Soner and Shreve [2], where

$$\begin{aligned} &dX = b(X)dt + \sigma(X)dW + dJ, \\ &J(t) = \int_0^t v(s)dK(s), \end{aligned} \tag{6.1}$$

where $|v(s)| \equiv 1$, and $K(\cdot)$ is a real-valued, right continuous, and nondecreasing singular control with $K(0) = 0$. Thus the control consists of the direction vectors $v(s)$ and the integral of the "force." For $\beta > 0$, the cost is

$$V(X(0), J, W) = E\int_0^\infty e^{-\beta t} k(X(t))dt + E\int_0^\infty e^{-\beta t} dK(t), \tag{6.2}$$

for a bounded and continuous $k(\cdot)$. For ease of development, a two-dimensional case will be used. It should be clear, however, that the method works in any dimension.

For computational reasons, $X(\cdot)$ must be confined to a bounded set G. For concreteness, we use the box $G = [-g^1, g^1] \times [-g^2, g^2]$ where $g^i > 0$, and we reflect orthogonally to the boundary when on the boundary. Thus the *computational model* is

$$X(t) = X(0) + \int_0^t b(X(s))ds + \int_0^t \sigma(X(s))dW(s) + J(t) + Y(t) - U(t), \tag{6.3}$$

where $Y(\cdot)$ and $U(\cdot)$ are the reflection terms; the $Y^i(\cdot)$ (respectively, $U^i(\cdot)$) can increase only when $X^i(t) = -g^i$ (g^i, respectively). Again, we use a rectangular grid G_h in G and define the extended grid G_h^+ and boundaries ∂G_h^+ and ∂G_h analogously to the definitions in §3.

For the choice of transition probabilities for the chain, we have the same three cases as in §3. The diffusion step is handled exactly as was Case 1 (§3). When $x = \xi_n^h \in \partial G_h^+$, we reflect to the closest point on ∂G_h, so the reflection step is simpler here, and no randomization is needed.

The control step (Case 2 of §3) is somewhat different. Refer to Fig. 4, where $\xi_n^h = x$, and the next step is to be a control step. The form of the original model (6.1) implies that the state increment in the control step can be in any direction and with any magnitude. For programming simplicity for the Markov chain model, we limit the

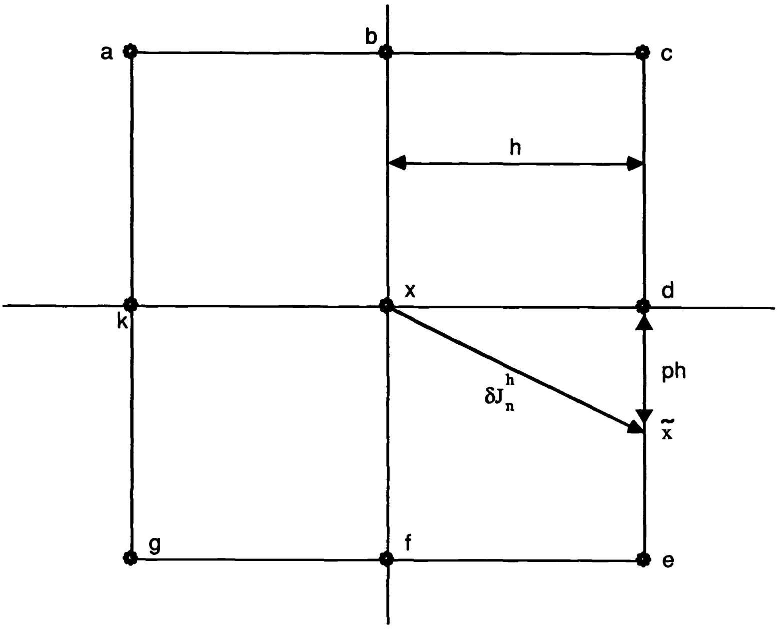


FIG. 4. *The control directions.*

magnitude of the control increments so that the state transitions are local. To realize an arbitrary direction, randomizations are used as in §3. In particular, we suppose for the Markov chain model that the control actions are the vectors connecting x to points on the boundary of the square in the figure. Each point on the boundary defines a direction and a magnitude of the control action. This "size" restriction on the magnitude can be removed. From the theoretical point of view, it is not needed. Let the δJ_n^h in the figure be the actual desired control action. This action takes x to a point $\tilde{x}$ on the boundary of the square. As in §3, $\tilde{x}$ is not usually a grid point, and we define the transition probability by taking δJ_n^h to be the *conditional mean value* $(E_n^h \delta\xi_n^h)$ of the increment in the state. Analogously to the situation in §3, for this example the control step transition probabilities are obtained by the randomization

$$p^h(x, d|\delta J_n^h) = (1 - p) = 1 - p^h(x, e|\delta J_n^h), \tag{6.4}$$

where p is defined in the figure. For consistency with (6.2), the cost associated with this increment δJ_n^h is the "magnitude" $\delta K_n^h \equiv |\delta J_n^h| = h(1 + p^2)^{1/2}$. Define the direction v_n^h by $\delta J_n^h = v_n^h \delta K_n^h$.

Let $p^h(x, y|\delta J)$ denote the transition probabilities under control action δJ. If no control is used at a step, we write the transition probability simply as $p^h(x, y)$. Then the dynamic programming equation analogous to (3.12), (3.13) is: For $x \in \partial G_h^+$,

$$V^h(x) = V^h(\text{nearest point to } x \text{ on } \partial G_h); \tag{6.5}$$

for $x \in G_h$,

$$\begin{aligned} V^h(x) = \min \Big\{ & e^{-\beta \Delta t^h(x)} \textstyle\sum_y p^h(x, y) V^h(y) + k(x) \Delta t^h(x), \\ & \min_{\delta J} \Big[\textstyle\sum_y p^h(x, y|\delta J) V^h(y) + |\delta J| \Big] \Big\}. \end{aligned} \tag{6.6}$$

The inner minimum concerns the minimum cost over all possible control actions, given that we use some control. The outer minimum chooses the better of (no control, the best control).

6.1. The inner minimum in (6.6). The computation of the inner minimum in (6.6) is not as formidable as it might seem. There are eight segments to consider, depending on where δJ points, namely, $(a,b),\cdots,(g,k)$, (k,a). We take the minimum over each segment separately. Consider segment (d,e). Then the minimum over the segment is (where p is defined in Fig. 4)

$$\min_{0\leq p\leq 1}[(1-p)V^h(x+e_1h)+pV^h(x+e_1h-e_2h)+h(1+p^2)^{1/2}]. \tag{6.7}$$

If $V^h(x+e_1h-e_2h)\geq V^h(x+e_1h+e_2h)$, then segment (d,e) will never be preferred to segment (c,d). In this way, we reduce the number of segments that need to be considered to four.

To evaluate the minimum in (6.7), note that if $V^h(x+e_1h)\leq V^h(x+e_1h-e_2h)$, then zero is the best value of p. If $[V^h(x+e_1h-e_2h)-V^h(x+e_1h)]/h\leq -1/\sqrt{2}$, then the best value of p is unity. Otherwise, the best value is obtained by differentiating and satisfies

$$\frac{-p}{\sqrt{1+p^2}}=\frac{[V^h(x+e_1h-e_2h)-V^h(x+e_1h)]}{h}.$$

The above procedure could be simplified by using only a finite number of possible values for the δJ. With appropriate choices, the optimal cost functions can be approximated as well as desired.

6.2. The convergence $\mathbf{V^h(x)\to V(x)}$. Define

$$J_n^h=\sum_{i=0}^{n-1}\delta J_n^h, K_n^h=\sum_{i=0}^{n-1}\delta K_i^h.$$

Define $\xi^h(\cdot)$, $J^h(\cdot)$, $K^h(\cdot)$, and the scaled processes $\hat{\xi}^h(\cdot),\cdots$, analogously to what was done in §4. Then Theorems 5.3–5.6 continue to hold with the following modifications. For Theorem 5.3, $\hat{J}^h(\cdot)\Rightarrow\hat{J}(\cdot)$, $\hat{K}^h(\cdot)\Rightarrow\hat{K}(\cdot)$, but $\hat{J}(\cdot)$ has the representation $\hat{J}(t)=\int_0^t\hat{v}(s)d\hat{K}_0(s)$, for a $\hat{v}(\cdot)$ and $\hat{K}_0(\cdot)$ satisfying $|\hat{v}(s)|\equiv 1$ and $\hat{K}_0(t)\leq\hat{K}(t)$ for all t. Also, for $\xi_0^h=x$,

$$\begin{aligned}\xi(t)=x&+\textstyle\int_0^t B(\xi(s))ds+\int_0^t\sigma(\xi(s))dW(s)\\&+\textstyle\int_0^t v(s)dK_0(s)+Y(t)-U(t).\end{aligned} \tag{6.8}$$

The fact that $\hat{K}_0(t)$ might be less than $\hat{K}(t)$ (or, equivalently, $K_0(t)\leq K(t)$) for some t is due to the fact that rapid time variations in the directions $\hat{v}^h(s)$ (the continuous parameter interpolation of the $\{v_n^h\}$ in the "stretched out" timescale) can cause a "cancellation" or "reduction" in the effects of the control. For example, consider the case where $\delta J_{n+1}^h=-\delta J_n^h$, for a consecutive sequence of values of n. Then the effects of this sequence would not appear in the limits $\hat{J}(\cdot)$ or $J(\cdot)$ since these limits are "averages" of the directions (loosely speaking), but they would appear in $\hat{K}(\cdot)$ and $K(\cdot)$.

The analogue of Theorem 5.4 holds, but with

$$\underline{\lim}_{h} V^h(x, J^h) \geq V(x, J, W).$$

Theorem 5.5 holds and so does Theorem 5.6. The proof of the analogue of Theorem 5.6 involves first showing that for any $\delta > 0$ there is a simple form of a δ-optimal control that has only finitely many impulses, each being bounded and taking a discrete number of values. Then the argument of Theorem 5.6 can be carried over.

7. A constrained and ergodic singular control problem. Ergodic singular control problems require some variations of the methods of the previous sections. To illustrate the general ideas, we will work with a particular but important problem class of the type developed by Harrison [8], Wein [9], [13], and Harrison and Wein [12] to model the input control for a queueing network with several customer classes with priorities and "throughput" constraints. The details of the development and analysis of the models are in the references and will only be briefly discussed, so that we can concentrate on the numerics. A numerical study of the model was carried out in [13] using the Markov chain approximation method, but no convergence proof was available. Our notation will be different from that in the references. Because of the constraint and the absence of a cost on the use of singular control, the model has some additional interest. The technique readily specializes to the more standard "ergodic" situations.

The basic system model is

$$X(t) = X(0) + bt + \sigma W(t) + J(t), \quad X(t) \in R^r, \tag{7.1}$$

where $J(\cdot)$ is a control that takes the special form

$$J^i(t) = K^i(t) - K^{r+1}(t), \quad i \leq r, \tag{7.2}$$

where the $K^i(\cdot)$ are the singular controls, i.e., they are right continuous, nondecreasing, and satisfy $K^i(0) = 0$. The $K(\cdot) = (K^1(\cdot), \cdots, K^{r+1}(\cdot))$ is said to be *admissible* or $(K(\cdot), W(\cdot))$ is said to be an *admissible pair* if $K(\cdot)$ is nonanticipative with respect to $W(\cdot)$. The results can readily be extended to include state-dependent drift and covariance, but for simplicity we concentrate on the cited problem. The system is subject to the constraints

$$\overline{\lim_{T}} \frac{1}{T} EK^i(t) \leq \lambda_i, \quad i = 1, \cdots, r+1, \quad \lambda_i > 0, \tag{7.3}$$

and the cost is

$$\gamma(K) = \overline{\lim_{T}} \frac{1}{T} E \int_0^T k(X(t))dt,$$

for a continuous $k(\cdot)$. Write $\lambda = (\lambda_1, \cdots, \lambda_{r+1})$ and define $\bar{\gamma}(\lambda) = \inf \gamma(K)$, where the inf is over all admissible controls such that (7.3) holds and $X(\cdot)$ is stationary. The λ-dependence of $\bar{\gamma}(\cdot)$ will be needed in §10.

The physical model leading to (7.1)–(7.3) contains $r+1$ servers, and $K^i(\cdot)$ is the heavy traffic limit of the suitably scaled, idle time of server i. The $X^i(\cdot)$ is called a workload imbalance in [8], [9], [13]. Roughly speaking, the workload of server i at time t is the (suitably scaled) amount of work that exists for server i from all the customers in the system at time t. The workload imbalance is a weighted difference of the workload of servers $i \leq r$ and $r+1$. This formulation is a very clever way of simplifying the problem when there are many classes of customers with priorities. The constraint (7.3) guarantees a maximum throughput for the system. Although the problem has a special structure, "ergodic" forms of the problems in the previous sections can be dealt with by the same techniques.

7.1. The "numerical" state space. For simplicity in the development, we work with $r = 2$. The method will obviously work for any r. The procedures and results of the previous sections will be used where possible. For numerical purposes, we need to bound the state space and use a setup somewhat similar to that in §3. Define g^i and the "box" G as in §6. The process $X(\cdot)$ will be confined to G for numerical purposes. Normally in such problems, we need to experiment with the "numerical boundary" to find one that is a suitable compromise between computational efficiency and minimal interference with the essential features of the optimal value function and control. For numerical purposes, we let $X(\cdot)$ reflect instantaneously on ∂G, with the reflection directions being orthogonal to the boundary except at the corners, where they point "diagonally" in.

With this modification (7.1) becomes

$$X(t) = X(0) + bt + \sigma W(t) + J(t) + Y(t) - U(t). \tag{7.4}$$

$Y(\cdot)$ and $U(\cdot)$ are the usual reflection terms. They are nondecreasing, take value zero at $t = 0$, and $Y^i(\cdot)$ (respectively, $U^i(\cdot)$) can increase only at points t for which $X^i(t) = -g^i$ (respectively, g^i). Those properties uniquely characterize $Y(\cdot)$ and $U(\cdot)$ (Theorem 5.1). For the problem in [9], [13], $Y(t) = U(t) = 0$, for all t, if G is large enough, since the optimal $J(\cdot)$ confines the $X(\cdot)$ to a bounded set.

The computational problem for even noncontrolled ergodic problems is still in its infancy, also. One promising approach to the computation of invariant measures is in [19].

8. The Markov chain approximation. The procedure of §3 will be followed to construct an approximating chain and control problem. Again, the numerical procedure consists in solving the optimal control problem for the Markov chain model. The convergence proofs in §10 show that the optimal value functions for the chain converge to $\bar{\gamma}(\beta)$.

For the approximation parameter h, let the g^i be integral multiples of h. Let G_h and G_h^+ denote the h-grid and extended h-grid, respectively, on G, as in §3. We also use the other notation of §3, where applicable. The types of state transitions are divided into the same three cases as in §3. The reflection case (Case 3) (for $\xi_n^h \in \partial G_h^+$) is trivial here, since we simply reflect back to the nearest point on ∂G_h. The diffusion step is also as for Case 1, and any transition function $p^h(x,y)$ can be used provided that (3.1) holds. Indeed, due to the lack of state dependence here, it is easier to construct transition functions with the required properties. Also $\Delta t^h(x) = \Delta t^h$, not depending on x. In general (3.4) implies that $\Delta t^h = h^2/Q$ for some real $Q > 0$.

8.1. The control step. The transition function for the control step in §§3 or 6 was obtained by first choosing the mean direction and then randomizing. Due to the particular structure of $J(\cdot)$ here, the randomization is not necessary, although it might be in problems where $J(\cdot)$ has a different structure.

If step n is a control step for the chain, then set $\delta\xi_n^h = \delta J_n^h = (\delta J_n^{h,1}, \cdots, \delta J_n^{h,r})$, and write $\delta J_n^{h,i} = \delta K_n^{h,i} - \delta K_n^{h,r+1}$. Write $\delta K_n^h = (\delta K_n^{h,1}, \cdots, \delta K_n^{h,r+1})$. We use $r = 2$

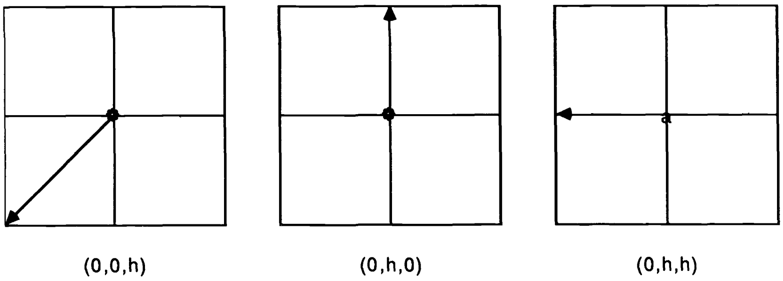


FIG. 5. *Examples of control directions for the ergodic problem.*

until further notice. If n is not a control step, then define $\delta J_n^h = \delta K_n^h = 0$. Suppose that $\xi_n^h = x \in G_h$, and we elect to exercise control. Then, as in §3, the transition function $p^h(x, y|\text{control})$ must be "locally" consistent with (7.4). We let the values of the $\delta K_n^{h,i}$ be either 0 or h. We could allow the increments to take values in any bounded interval $[0, k_0]$ or in the set of points $[\ell h,\ \ell = 0, \cdots, \ell_0,\ \ell_0 < \infty]$ and still get the same convergence results. It is usually simpler, however, for the programming to work with "local transitions." For our case $r = 2$, there are seven possibilities for the control actions (we eliminate the cases $\delta K = (0,0,0)$ and (h,h,h), since they keep us at the same state). Write $x = (x^1, x^2)$. Then the possible control actions are (refer to Fig. 5) $\delta K = (0,0,h)$, yielding $\xi_{n+1}^h = (x^1 - h, x^2 - h)$; $\delta K = (0,h,0)$, yielding $\xi_{n+1}^h = (x^1, x^2 + h)$; $\delta K = (h,0,0)$, yielding $\xi_{n+1}^h = (x^1 + h, x^2)$; $\delta K = (0,h,h)$, yielding $\xi_{n+1}^h = (x^1 - h, x^2)$; $\delta K = (h,0,h)$, yielding $\xi_{n+1}^h = (x^1, x^2 - h)$; and $\delta K = (h,h,0)$, yielding $\xi_{n+1}^h = (x^1 + h, x^2 + h)$. The transition function for the case of general r should be obvious.

8.2. The control problem for the Markov chain. In [11, §11] a convergence theory for the discretization of ergodic control problems is given. Owing to both the constraint (7.3) and the "tightness problems" stemming from the nature of the singular control problem as discussed in §4, that development cannot be used directly here. We will develop the appropriate adaptation of the technique of §§3–6. Set $\Delta t_n^h = 0$ if step n is either a control or reflection step, and set $\Delta t_n^h = \Delta t^h$ otherwise. (In [13], $\Delta t_n^h = \Delta t^h$ was used for all n. The limit results are the same for both cases, since it can be shown that the "fraction of the number of steps" spent either on the boundary or controlling goes to zero as $h \to 0$.) Define the interpolation $\xi^h(\cdot)$ as in §4. For the Markov chain model, we will work with feedback policies and stationary chains. Since the state space is finite, for each feedback policy, there is some stationary chain. $\{\xi_n^h,\ n < \infty\}$ will refer to such a stationary chain. If the stationary chain under the chosen feedback policy is not unique, then pick any one. The choice will be irrelevant. Let $\{\delta K_n^h\}$ denote the sequence of control actions for the chain. We use K^h to denote the control policy. Define the interpolation $K^h(\cdot)$ (intervals $\{\Delta t_n^h\}$) from $\{\delta K_n^h\}$ analogously to (4.1).

Let $N^h(T) = \min\{n : t_n^h \geq T\}$. A literal translation of the continuous parameter problem into the stationary chain yields the cost

$$\lim_T \frac{E \sum_{n=0}^{N^h(T)-1} k(\xi_n^h) \Delta t_n^h}{T} \tag{8.1}$$

and the constraint

$$\lim_T \frac{E\sum_{n=0}^{N^h(T)-1} \delta K_n^{h,i}}{T} \le \lambda_i. \tag{8.2}$$

It is implied by the proof of Theorem 10.1 and the comments below (9.1) that if (8.2) holds, then the fraction of steps that are either control or reflection steps is $O(h)$. By using this, (8.1) and (8.2) can be approximated by (with errors of the order $O(h)$ in (8.4))

$$\lim_N \frac{1}{N} E \sum_{n=0}^{N-1} k(\xi_n^h) \equiv \gamma^h(K^h), \tag{8.3}$$

$$\lim_N \frac{1}{N} E \sum_{n=0}^{N-1} \delta K_n^{h,i} \le \lambda_i \Delta t^h. \tag{8.4}$$

For the same reasons, the left sides of (8.1) and (8.2) equal (modulo $O(h)$), respectively,

$$\lim_T \frac{1}{T} \int_0^T Ek(\xi^h(s))ds, \qquad \lim_T \frac{1}{T} EK^{h,i}(T).$$

Define $\bar{\gamma}^h(\lambda) = \inf_{K^h}(K^h)$, subject to (8.3) and the stationarity of $\{\xi_n^h,\ n < \infty\}$.

Due to the constraint (8.2), a dynamic programming formulation such as (3.12) and (3.13) cannot be used, and the optimization problem must be phrased in terms of a linear programming (LP) formulation [20], [21].

9. The LP formulation.

A control policy for $\{\xi_n^h,\ n < \infty\}$ is said to be *pure Markov* if for each state a unique control action is assigned. A control policy is said to be *randomized Markov* if for each state a distribution (depending only on the present state) of control actions is assigned. Due to the constraint, there might not be an optimal control in the class of pure Markov policies, but one always exists in the class of randomized Markov policies.

The LP formulation uses stationary chains $\{\xi_n^h,\ n < \infty\}$ and randomized Markov policies. Let $p^h(x, y \mid \delta K)$ denote the transition function, given a control action δK. We use $\delta K = 0 \equiv (0,0,0)$ to denote that no control is used. We use π^h to denote the stationary probabilities for the chain, i.e., $\pi^h(x) = P\{\xi_n^h = x\}$, $\pi^h(x, \delta K) = P\{\xi_n^h = x,\ \delta K_n^h = \delta K\}$, $\pi^h(\delta K^i = 0) = P\{\delta K_n^i = 0\}$, $\pi^h(G_h) = \sum_{x \in G_h} \pi^h(x)$, etc. The control to be used at any state is determined by the conditional distribution $P\{\delta K_n^h = \delta K \mid \xi_n^h = x\} = \pi^h(x, \delta K)/\pi^h(x)$.

Thus $\pi^h(x) = \sum_{\delta K} \pi^h(x, \delta K)$ and $\pi^h(\delta K) = \sum_x \pi^h(x, \delta K)$. The fact that the probabilities sum to unity yields the constraints

$$\pi^h(x) = \sum_{y,\delta K} p^h(y, x \mid \delta K)\pi^h(y, \delta K), \quad \sum_x \pi^h(x) = 1, \quad \pi^h(x, \delta K) \ge 0. \tag{9.1}$$

Following the development in [13], and using the fact that $\Delta t_n^h = 0$ if $\xi_n^h \in \partial G_h^+$ or if n is a control step, the constraint (8.2) can be written as

$$\frac{E\delta K_n^i}{E\Delta t_n^h} = \frac{h\pi^h(\delta K^i \ne 0)}{(h^2/Q)\pi^h(G_h, \delta K = 0)} \le \lambda_i$$

or

$$\pi^h(\delta K^i = h) \le \frac{\lambda_i h}{Q}\pi^h(G_h, \delta K = 0) = O(h). \tag{9.2}$$

Note that (9.2) implies the probability that any step n is a control step is $O(h)$. Using this and the fact that a reflection step must be followed by either a control or a diffusion step yields than the mean value of the total variation of the $U^h(\cdot)$ and $Y^h(\cdot)$ terms on the interpolated time interval $[0,T]$ is (modulo $O(h)$ times the value of (9.3))

$$\pi^h(\partial G_h^+) \cdot h(T/\Delta t^h) = O(1/h)\pi^h(\partial G_h^+). \tag{9.3}$$

The proof of Theorem 10.1 below implies that the mean variation of the reflection terms for a stationary process satisfying the constraints is bounded uniformly in h on any fixed interpolated time interval. Thus by (9.3), $\pi^h(\partial G_h^+) = O(h)$. Using this and (9.2) yields that $\pi^h(G_h, \delta K = 0) = 1 - O(h)$, and (modulo $O(h^2)$) we can rewrite (9.2) as

$$\pi^h(\delta K^i = h) \le \lambda_i h/Q, \qquad i \le r+1. \tag{9.4}$$

An analogous argument applied to (8.1) yields the cost function

$$\frac{\sum_{x\in G_h} k(x)\pi^h(x, \delta K = 0)}{\pi(G_h, \delta K = 0)} = \sum_{x\in G_h} k(x)\pi^h(x) \equiv \gamma^h(K^h), \tag{9.5}$$

where the equalities are all modulo $O(h)$. Define $\gamma^h = \inf_{K^h}\gamma^h(K^h)$, where the infimum is over the randomized Markov policies satisfying (9.4).

9.1. The LP problem. For the numerical procedure, the LP formulation of the Markov chain control problem is the minimization of the right-hand side of (9.5) subject to the constraints (9.1) and (9.4). The "activity" variables of the LP formulation are the $\pi^h(x, \delta K)$. The number of constraints totals $\bar{N}^h =$ (number of points in the state space G_h^+ plus 1) plus $(r+1)$, the first set of (9.1) yielding the first group and the second group coming from the $r+1$ constraints (9.4). $\bar{N}^h$ is the maximum number of nonzero variables in any basic solution to the LP. Thus there are at most $(r+2)$ states x that $\pi^h(x, \delta K)$ might be positive for more than one value of δK. Thus, there are at most $(r+2)$ states at which the control is actually "randomized," for any h. The effects of these states disappears as $h \to 0$.

Since the state space G_h^+ is finite, for each h and each randomized or pure Markov control policy, there is at least one stationary chain for each transition function. The LP solution gives a least cost stationary chain, whether or not the least cost chain is unique.

10. The convergence theorem.

10.1. Preliminary calculations. The following facts and definitions will be useful in Theorem 10.2. Let $\pi^h(\cdot)$ denote the optimal steady-state probabilities, via the LP (9.1), (9.4), (9.5). Let $\{\xi_n^h,\ \delta K_n^h,\ n < \infty\}$ denote the associated stationary chain and control policy. Let γ^h denote the associated ergodic cost. We have that $\delta\xi_n^h$ equals δY_n^h or δU_n^h if n is a reflection step and $\delta\xi_n^h = \delta K_n^h$ if n is a control step. If n is a "diffusion" step, then (see (3.4))

$$\delta\xi_n^h = b\Delta t^h + \beta_n + O(h^\rho \Delta t^h),$$

where

$$E_n^h \beta_n^h \beta_n^{h'} = \sigma\sigma' \Delta t^h + O(h^p \Delta t^h), \qquad E_n^h \beta_n^h = 0.$$

Thus

$$\xi_{n+1}^h = \xi_0^h + bn\Delta t^h + \sum_{i=0}^{n} \beta_i^h + J_{n+1}^h + Y_{n+1}^h - U_{n+1}^h + (\text{negligible terms}).$$

The negligible terms are

$$\sum_{i=0}^{n-1} O(h^p \Delta t_i^h) - b\Delta t^h (\text{number of reflection or control steps in } [0, n]).$$

The continuous parameter interpolation is (see §4 for the definitions)

$$\xi^h(t) = \xi^h(0) + bt + M^h(t) + J^h(t) + Y^h(t) - U^h(t) + (\text{negligible terms}). \tag{10.1}$$

Let $0 < t_1 < \cdots < t_q$, where $|t_{i+1} - t_i| > \Delta t^h$ for small h. Then due to the stationarity of $\{\xi_n^h,\ n < \infty\}$ and the independence of Δt^h of x, the distribution of $\{\xi^h(t_0 + t_i),\ K^h(t_0 + t_i) - K^h(t_0),\ i \le q\}$ does not depend on t_0 for small h. Also, by the stationarity and the constraint (9.4),

$$E[K^{h,i}(t_0 + t) - K^{h,i}(t_0)] \le \lambda_i t + O(h) \quad \text{for all } t, t_0. \tag{10.2}$$

The problem of proving tightness for $\{K^h(\cdot)\}$ is the same here as that in §4, and we cannot necessarily find a weakly convergent subsequence of $\{\xi^h(\cdot),\ K^h(\cdot), \cdots\}$ in the Skorokhod topology. However, a variant of the rescaling method used in Theorem 5.3 will work. Define the processes $\hat{T}^h(\cdot)$, $\hat{\xi}^h(\cdot)$, $\hat{K}^h(\cdot), \cdots$, as in §4 and Theorem 5.3.

The "tightness" situation is better here than in the problem of §§2–6, since the probability that any n is a control step is $O(h)$. It will be seen that this implies that $\hat{T}(\cdot)$ and $\xi(\cdot)$ are continuous w.p.1 at each t. In the convergence proof we will need to show that $\xi^h(t)$ converges to $\xi(t)$ in distribution for each t. A basic tool is formally stated in the following "continuous mapping" lemma, whose proof is easy and is omitted. Let R denote the real line.

LEMMA 10.1. *Let $\hat{\tau}(\cdot)$ be a nonnegative, nondecreasing function with $\hat{\tau}(0) = 0$, which is right continuous and let $\hat{\tau}(t) \to \infty$ as $t \to \infty$. Define $\tau(\cdot)$ by $\tau(t) = \inf\{s : \hat{\tau}(s) > t\}$. Then $\tau(\cdot) \in D[0, \infty)$. Define the map $\Phi : D[0, \infty) \to R$ by $\Phi(\hat{\tau}(\cdot)) = \tau(t_0)$. Suppose that $\tau(\cdot)$ is continuous at t_0. Then $\Phi(\cdot)$ is continuous at $\hat{\tau}(\cdot)$.*

10.2. The convergence theorem: Part 1.

THEOREM 10.2. *The sequence $\{\hat{\xi}^h(\cdot), \cdots, \hat{T}^h(\cdot)\}$ is tight. Abusing notation, let h index a weakly convergent subsequence with limit denoted by $(\hat{\xi}(\cdot),\ \hat{K}(\cdot), \cdots, \hat{T}(\cdot))$. Define $T(\cdot)$, $\xi(\cdot)$, $K(\cdot), \cdots$ as in Theorem 5.3. Then there is a standard vector-valued Wiener process $W(\cdot)$ such that*

$$\xi(t) = \xi(0) + bt + \sigma W(t) + J(t) + Y(t) - U(t), \tag{10.3}$$

where $J^i = K^i - K^{r+1}$ and $K(\cdot)$ is nonanticipative with respect to $W(\cdot)$. The $Y(\cdot)$ and $U(\cdot)$ are the appropriate reflection terms. The distribution of $(\xi(t_0 + \cdot),\ K(t_0 + \cdot) - K(t_0))$ does not depend on t_0. Also,

$$\begin{gathered} \gamma^h = \gamma^h(K^h) \to \gamma(K) = E\int_0^T k(\xi(s))ds/T, \quad \text{for each } T > 0, \\ EK^i(t) \le \lambda_i t, \quad \text{for all } t. \end{gathered} \tag{10.4}$$

Proof. The tightness of $\{\hat{\xi}^h(\cdot), \cdots, \hat{T}^h(\cdot)\}$ is proved as in Theorem 5.3, as is representation (10.3) and the nonanticipativness of $K(\cdot)$. For any $t, s \geq 0$, (10.2) implies that

$$E|K^i(t+s) - K^i(t)| \leq \overline{\lim_h} E|K^{h,i}(t+s) - K^{h,i}(t)| \leq \beta_i s. \tag{10.5}$$

The bound on $\{K^{h,i}(\cdot)\}$ in (10.2) implies that $\hat{T}(t) \to \infty$ w.p.1 as $t \to \infty$. Let $\mathcal{T}$ denote the set $\mathcal{T} = \{t: P\{T(\cdot)$ is discontinuous at $t\} > 0\}$. We claim that $\mathcal{T}$ is empty. The proof is as follows: Let $T(0^-) = 0$. Given $t_0 \geq 0$, suppose that there are $\delta_i > 0$ such that

$$P\{T(t_0^+) - T(t_0^-) \geq \delta_1\} \geq \delta_0.$$

This can only happen if for any $\delta_2 > 0$,

$$\overline{\lim_h} P\{|K^h(t_0 + \delta_2) - K^h(t_0 - \delta_2)| \geq \delta_1/2\} \geq \delta_0/2.$$

But due to the arbitrariness of δ_2, this latter inequality is impossible by (10.5). This implies that the probability is zero that $T(\cdot)$ will jump more than δ_1 at t_0, which proves the claim, since δ_1 is arbitrary. Thus $T(\cdot)$ is continuous w.p.1 at each $t \geq 0$, and so is $K(\cdot)$. Now by representation (10.3) and Theorem 5.1, $Y(\cdot)$, $U(\cdot)$, and $\xi(\cdot)$ are also continuous w.p.1 at each $t \geq 0$.

Now define the inverse function $T^h(t) = \inf\{s: \hat{T}^h(s) > t\}$. We can write

$$\xi^h(t) = \hat{\xi}^h(T^h(t)), \qquad K^h(t) = \hat{K}^h(T^h(t)). \tag{10.6}$$

Let q be an integer and let $0 \leq t_1 < \cdots < t_q$. Since $T(\cdot)$ is continuous w.p.1 at each t, we have that $\{T^h(t_i), i \leq q\}$ converges weakly to $\{T(t_i), i \leq q\}$. Now Lemma 10.1, the definitions of $T^h(\cdot)$ and $T(\cdot)$, the weak convergence of $\{\hat{\xi}^h(\cdot), \cdots, \hat{T}^h(\cdot)\}$ to $(\hat{\xi}(\cdot), \cdots, \hat{T}(\cdot))$, the w.p.1 continuity of $\hat{\xi}(\cdot)$ and $\hat{K}(\cdot)$ at each t, and (10.6) yield (in the sense of weak convergence)

$$\xi^h(t_i) \to \hat{\xi}(T(t_i)) = \xi(t_i), \qquad i \leq q, \tag{10.7}$$

$$K^h(t_{i+1}) - K^h(t_i) \to \hat{K}(T(t_{i+1})) - \hat{K}(T(t_i)) = K(t_{i+1}) - K(t_i), \qquad i \leq q.$$

By (10.7) and the fact that the distribution of $(\xi^h(t_0 + \cdot), K^h(t_0 + \cdot) - K^h(t_0))$ does not depend on t_0 (in the sense used above (10.2)), we have that $\xi(\cdot)$ is stationary and the distribution of $(\xi(t_0 + \cdot), K(t_0 + \cdot) - K(t_0))$ does not depend on t_0. By the stationarity properties, the cost can be represented as follows:

$$\begin{aligned} t\gamma^h &= E\int_0^t k(\xi^h(s))ds + O(h) \\ &= E\int_0^{T^h(t)} k(\hat{\xi}^h(s))d\hat{T}^h(s) + O(h) \\ &\to E\int_0^{T(t)} k(\hat{\xi}(s))d\hat{T}(s) = E\int_0^t k(\xi(s))ds \\ &= tEk(\xi(0)) = t\gamma(K), \end{aligned}$$

and the theorem is proved. □

Theorem 5.2 implies that

$$\underline{\lim_h} \bar{\gamma}^h(\lambda) \geq \bar{\gamma}(\lambda).$$

10.3. The convergence theorem, completed. For the problem of §2, it was not hard to find a convenient δ-optimal comparison control so that the "reverse inequality," $\overline{\lim}_h V^h(x) \le V(x)$ could be obtained in Theorem 5.6. It is more difficult to choose a "comparison control" for the ergodic problem. We need to find a "nice" δ-optimal control for the ergodic problem for (7.4) with constraint (7.3), which can be adapted to $\{\xi_n^h, n < \infty\}$ so that a result such as in Theorem 5.6 can be proved. Little is known concerning optimal or δ-optimal controls for singular control problems for either the ergodic or nonergodic case. To proceed, we need to make an assumption on the existence of a δ-optimal control of a particular form.

Numerical experience and the literature concerning problems where an optimal control has been characterized (for ergodic or not ergodic problems) [2], [9], suggest that Assumption 10.2, below, is quite reasonable. Assumption 10.1 is needed, since the constraint (9.4) is not necessarily satisfied when the control of Assumption 10.2 is applied to the chain, and it is necessary to perturb the constraint slightly. Assumption 10.1 also does not seem to be restrictive, although we have not been able to prove that it holds in general. A typical control boundary from [13] is shown in Fig. 6.

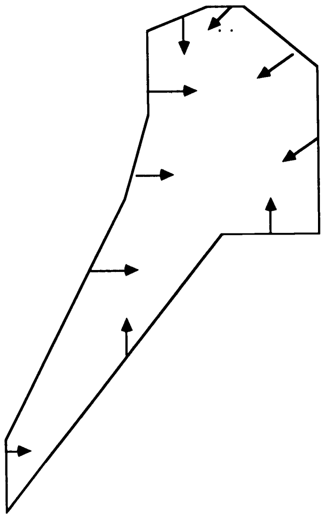


FIG. 6. *A numerical solution: Piecewise linear approximation to the boundary.*

Assumption 10.1. $\bar{\gamma}(\cdot)$ is continuous at λ if all $\lambda_i > 0$.

For $\delta_0 > 0$ and $\lambda_i - \lambda_0 > 0$, define $\lambda - \delta_0 = (\lambda_1 - \delta_0, \cdots, \lambda_{r+1} - \delta_0)$.

Assumption 10.2. For any small $\delta_0 > 0$ and constraint vector $\lambda - \delta_0$, there is a δ-optimal control $K_\delta(\cdot)$ (define $J_\delta(\cdot) = \{K_\delta^i(\cdot) - K_\delta^3(\cdot), i = 1,2\}$) satisfying the following conditions. There is a set $\hat{G}$ that is the closure of an open set in G and is such that the boundary $\partial\hat{G}$ is composed of a finite number of segments $\partial\hat{G}_1, \cdots$, each being continuously differentiable. The segments are nontangent at the corners where they meet. $J_\delta(\cdot)$ reflects to the "interior" of $\hat{G}$, and the directions of reflection on $\partial\hat{G}$, are constant on each $\partial\hat{G}_i$. The reflection directions $\{\bar{v}_i\}$ are taken from the set of seven directions defined in §7. The reflectings are not in opposite directions on adjacent boundary segments. There is a weak sense unique stationary process $X_\delta(\cdot)$

satisfying

$$X_\delta(t) = X_\delta(0) + bt + \sigma W(t) + J_\delta(t) \tag{10.8}$$

and

$$P\{X_\delta(t) \text{ is an } \varepsilon\text{-neighborhood of any “corner” of } \partial\hat{G}\} \xrightarrow{\varepsilon} 0. \tag{10.9}$$

Remark. By “reflecting to the interior,” we mean that the angles between $\bar{v}_i$ and the outward normals to $\partial\hat{G}_i$ are strictly greater than $\pi/2$.

Remark. The condition seems broad as it stands, since the boundary segments can be made aritrarily small. But we note that the condition can be weakened to allow continuous (rather than constant) reflection directions on each segment, subject to the “interior pointing” and “corner” conditions in Assumption 10.2.

THEOREM 10.3. *Assume that Assumptions* 10.1 *and* 10.2 *hold true. Then*

$$\bar{\gamma}^h(\lambda) \to \bar{\gamma}(\lambda).$$

Proof. We need only prove that

$$\overline{\lim_h} \bar{\gamma}^h(\lambda) \leq \bar{\gamma}(\lambda). \tag{10.10}$$

Let $\delta > 0$. By Assumption 10.1, choose small $\delta_0 > 0$ such that

$$\bar{\gamma}(\lambda - \delta_0) \leq \bar{\gamma}(\lambda) + \delta. \tag{10.11}$$

Following the general approach in Theorem 5.6, we “adapt” $J_\delta(\cdot)$ to $\{\xi_n^h,\ n < \infty\}$ as follows. The state space is just $\hat{G} \cap G_h$. For any step n where distance $(\xi_n^h, \partial\hat{G}) > h$, use the diffusion step transition probabilities. Otherwise, the step will be a control step with the increment $\delta\xi_n^h \equiv \delta J_n^h$ being in the reflection direction of the nearest boundary segment. If two boundary segments are equally close, use any one of the two directions. By (10.9), the choice will not matter in the limit. As in §8, we can write $\delta J_n^h = \delta K_n^i - \delta K_n^3$ for some $i = 1, 2$, where one or more of the components of $\{\delta K_n^i,\ i \leq 3\}$ are positive. Analogously to the notation in §8, let $K_\delta^h(\cdot)$ denote the “adapted” control policy and accordingly define $J_\delta^h(\cdot)$.

For each h, there is at least one stationary process under the given control policy. Let $\{\xi_n^h,\ n < \infty\}$ be such a stationary process and define $\xi^h(\cdot)$, $\hat{\xi}^h(\cdot)$, $\hat{T}^h(\cdot), \cdots$, in the usual way. Suppose for the moment that (it will be shown to hold below)

$$\sup_h E|K^h(t)|^2 < \infty, \quad \text{for each } t < \infty. \tag{10.12}$$

Then we have

$$P\{\text{control used at step } n\} = O(h). \tag{10.13}$$

(See (9.2) for a related calculation.) Using (10.13) and a proof similar to that of Theorem 10.2, it can be shown that $\gamma^h(K_\delta^h) \to \gamma(K_\delta)$ and that the limit process $\xi(\cdot)$ satisfies (10.8) with control $K_\delta(\cdot)$, and that the distributions of $(\xi(t_0 + \cdot),\ K_\delta(t_0 + \cdot) - K_\delta(t_0))$ do not depend on t_0.

As in Theorem 10.2, (10.13) implies that the set $\mathcal{T}$ (defined in Theorem 10.2) is empty. Thus, $K_\delta^h(t) \to K_\delta(t)$ in distribution for each $t \geq 0$. Hence, (10.12) implies

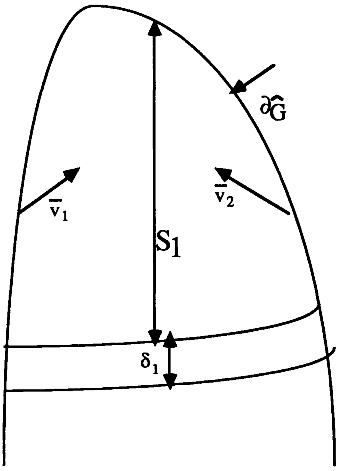


FIG. 7. *A boundary sector.*

that $EK^h_\delta(t) \to EK_\delta(t)$. Since $EK_\delta(t) \le (\lambda_i - \delta_0)t$, we have $EK^h_\delta(t) \le \lambda_i t$ for small h. Since $\bar{\gamma}^h(\lambda)$ is the optimal ergodic value for the stationary chain under constraint value β, we have that $\bar{\gamma}^h(\lambda) \le \gamma^h(K^h_\delta)$ for small h. Since, by Assumption 10.2, $K_\delta(\cdot)$ is δ-optimal under constraint vector $\lambda - \delta_0$, we have $\gamma(K_\delta) \le \bar{\gamma}(\lambda - \delta_0) + \delta$. The above facts imply that for small h, $\bar{\gamma}^h(\lambda) \le \gamma^h(K^h_\delta) \to \gamma(K_\delta) \le \bar{\gamma}(\lambda - \delta_0) + \delta$. Now (10.10) follows from this, (10.11), and the arbitrariness of δ and δ_0. Thus only (10.12) remains to be proved.

To prove (10.12), we adapt the proof of [7, Thm. 7]. Recall that $Y^h_n = U^h_n \equiv 0$ here, and write $\xi^h(t)$ in the form

$$\xi^h(t) = \xi^h(0) + D^h(t) + J^h_\delta(t), \tag{10.14}$$

where (10.14) defines $D^h(t)$. Let S_1 denote the intersection of a closed disc with $\hat{G}$ such that at most half of two adjacent segments of $\partial\hat{G}$ are included. For ease of description, refer to Fig. 7, where a "typical" S_1 is defined. Let $N_h(S)$ denote the h neighborhood of a set S. For small $\delta_1 > 0$, define the stopping times $\{\tau^h_n\}$ recursively by

$$\tau^h_1 = \min\{t \colon \xi^h(t) \in N_h(\partial\hat{G} \cap S_1)\} \wedge 1,$$

$$\tau^h_{2m} = \min\{t > \tau_{2m-1} \colon \text{distance}(\xi^h(t), S_1) \ge \delta_1\} \wedge 1,$$

$$\tau^h_{2m+1} = \min\{t > \tau_{2m} \colon \xi^h(t) \in N_h(\partial\hat{G} \cap S_1)\} \wedge 1.$$

Define $N^h = \min\{m \colon \tau^h_{2m} = 1\}$.

By (10.14), we can write

$$J^h_\delta(1) = \left(\sum_{m=1}^{N^h+1} [\xi^h(\tau^h_{2m}) - \xi^h(\tau^h_{2m-1})] - \sum_{m=1}^{N^h+1} [D^h(\tau^h_{2m}) - D^h(\tau^h_{2m-1})] \right). \tag{10.15}$$

By (10.15) and the properties of $D^h(\cdot)$, we see that there are $c_i > 0$ such that

$$E|J^h(1)|^2 \le c_1 + c_2 E|N^h|^2.$$

So we need only bound $E|N^h|^2$. Refer again to Fig. 7. There is $\delta_2 > 0$ such that (for small h) for $\xi^h(t)$ to go from the exterior of $N_{\delta_1}(S_1) \cap \hat{G}$ at some time t to $N_h(S_1 \cap \partial\hat{G})$ at some time $t+s$, we need $|D^h(t+s) - D^h(t)| \geq \delta_2$. Recall that $D^h(\cdot)$ is the interpolation of a sum of terms that are either proportional to Δt^h or else are martingale differences whose variances are $O(\Delta t^h)$. Thus given $\delta_3 > 0$, there is $\delta_4 > 0$ such that for all finite stopping times τ and small h,

$$(10.16) \qquad P\{\sup_{\tau+\delta_4 \geq s \geq \tau} |\, D^h(s) - D^h(\tau)| \geq \delta_2 |\text{ data up to time } \tau\} \leq 1 - \delta_3.$$

Inequality (10.16) implies that (for $m \geq 1$)

$$(10.17) \qquad P\{\tau^h_{2m+1} - \tau^h_{2m} \geq \delta_4 \mid \text{data up to time } \tau^h_{2m}\} \geq \delta_3.$$

But (10.17) implies that all moments of N^h are bounded uniformly in h. Thus $\sup_h E|J^h(1)|^2 < \infty$. Inequality (10.12) follows from that by the stationarity and the fact that the $\bar{v}_i$ on adjacent boundary segments do not point in opposite directions. □

REFERENCES

[1] M. Taksar, *Average optimal singular control and a related stopping problem*, Math. Oper. Res., 10 (1985), pp. 63–81.

[2] H. Mete Soner and S. E. Shreve, *Regularity of the value function for a two-dimensional singular stochastic control problem*, SIAM J. Control Optim., 27 (1989), pp. 876–907.

[3] J. P. Lehoczky and S. E. Shreve, *Absolutely continuous and singular stochastic control*, Stochastics, 17 (1986), pp. 91–110.

[4] P. L. Chow, J.-L. Menaldi, and M. Robin, *Additive control of stochastic linear systems with finite horizons*, SIAM J. Control Optim., 23 (1985), pp. 858–899.

[5] I. Karatzas and S. E. Shreve, *Equivalent models for finite fuel stochastic control*, Stochastics, 18 (1986), pp. 245–276.

[6] I. Karatzas, *A class of singular stochastic control problems*, Adv. Appl. Probab., 15 (1983), pp. 225–254.

[7] H. J. Kushner and L. F. Martins, *Routing and singular control for queueing networks in heavy traffic*, SIAM J. Control Optim., 28 (1990), pp. 1209-1233.

[8] J. M. Harrison, *Brownian models of queueing networks with heterogeneous customer populations*, in Stochastic Differential Systems, Stochastic Control Theory, and Applications, Fleming and Lions, eds., IMA, Vol. 10, Springer-Verlag, Berlin, 1988, pp. 147–186.

[9] L. M. Wein, *Optimal control of a two station Brownian network*, Math. Oper. Res., 15 (1990), pp. 215–242.

[10] H. J. Kushner, *Probability Methods for Approximations in Stochastic Control and for Elliptic Equations*, Academic Press, New York, 1977.

[11] ———, *Numerical methods for stochastic control in continuous time*, SIAM J. Control Optim., 28 (1990), pp. 999-1048.

[12] J. M. Harrison and L. M. Wein, *Scheduling networks of queues: Heavy traffic analysis of a two station closed network*, Oper. Res., to appear.

[13] L.M. Wein, *Scheduling networks of queues: heavy traffic analysis of a multistation network with controllable inputs*, Sloan School of Management, M.I.T., Boston, MA, preprint; Management Sci., to appear.

[14] P. Billingsley, *Convergence of Probability Measures*, John Wiley, New York, 1968.

[15] T.G. Kurtz, *Approximation of Population Processes*, Vol. 36, CBMS-NSF Regional Conference Series in Appl. Math., Society for Industrial and Applied Mathematics, Philadelphia, PA, 1981.

[16] M.I. Reiman, *Open queueing networks in heavy traffic*, Math. Oper. Res., 9 (1984), pp. 441–458.

[17] H.J. Kushner and K.M. Ramachandran, *Optimal and approximately optimal control policies for queues in heavy traffic*, SIAM J. Control Optim., 27 (1989), pp. 1293–1318.

[18] ———, *Nearly optimal singular controls for wideband noise driven systems*, SIAM J. Control Optim., 26 (1988), pp. 569–591.

[19] H.J. KUSHNER AND A.J. KLEINMAN, *Mathematical programming and the control of Markov chains*, Internat. J. Control, 13 (1971), pp. 801–820.

[20] C. DERMAN, *Denumerable state Markov decision processes—average criterion*, Ann. Math. Statist., 37 (1966), pp. 1545–1554.

[21] J. DAI, *Steady state analysis of reflected Brownian motions: characterization, numerical methods and queueing applications*, thesis, Operations Research Dept., Stanford University, Stanford, CA, 1990.

SIAM J. CONTROL AND OPTIMIZATION
Vol. 29, No. 6, pp. 1476–1508, November 1991

010

A FRAMEWORK FOR TWO-DIMENSIONAL HYPERSTABILITY THEORY BASED PROVABLY CONVERGENT ADAPTIVE TWO-DIMENSIONAL IIR FILTERING*

SANKAR BASU†

Abstract. A set of results for two-dimensional quarter plane causal systems reminiscent of one-dimensional hyperstability theory have been reported. The key to this development is a little known result of Landau [*Math. Ann.*, 62 (1906), p. 272], which asserts that a positive polynomial in two variables can be expressed as the sum of squares of polynomials in one variable whose coefficients are real rational functions of the other variable. The tools used are largely based on notions of passivity and the results obtained can be interpreted as a two-dimensional quarter plane causal generalization of the fact that if the total flow of energy into a dissipative system is upper bounded then both input and output asymptotically die out to zero. An adaptive two-dimensional recursive filtering scheme potentially useful in propagating wave type two-dimensional problems is considered next. It is then shown via our two-dimensional hyperstability results that the adaptive scheme converges in an appropriate sense.

Key words. two-dimensional system, passive scattering systems, adaptive filtering, circuit synthesis

AMS(MOS) subject classifications. 93C55, 93D15, 49E20

1. Introduction. The purpose of this paper is two-fold. The first is to develop two-dimensional (2-D) counterparts of some central aspects of the well-known hyperstability theory [1], [4], and the second is to show its applications in 2-D adaptive signal processing schemes mostly paralleling the one-dimensional developments reported in [28], [32], [36]. While our developments are largely motivated by potential applications in adaptive filtering, in view of the important role of hyperstability theory in various areas of system theory such as in control system systhesis [1], stochastic realization [2], digital filter stability [3], etc., we believe that 2-D extension of such a theory is of independent interest. In relating these apparantly disparate notions, the one-dimensional Kalman–Popov–Yakubovitch (KPY) lemma (otherwise known as the positive real lemma [5]), which provides a characterization of the property of dissipativeness of arbitrary (minimal) realization of systems, proves to be pivotal. While the KPY lemma can be derived via techniques akin to those known in linear quadratic optimal control theory [6], it can alternatively be viewed [5] as a consequence of synthesizability of positive (real) transfer functions or spectral factorability of parahermitian positive definite transfer matrices [2], or as a combination of both. Note that in the case of scalar transfer functions this last result is a reformulation of the fact that positive polynomials can be expressed as the sum of squares of two real polynomials. It is, however, well known [15] that this latter fact does not extend to 2-D dimensions in general. Our starting point in this work is a little known theorem of Landau [7] as a follow up of earlier work on the theory of curves by Hilbert [8] that positive polynomials in two variables can, in fact, be expressed as a sum of squares of polynomials in one variable whose coefficients are real rational functions of the other

* Received by the editors September 18, 1989; accepted for publication (in revised form) September 10, 1990.

† Department of Electrical Engineering and Computer Science, Stevens Institue of Technology, Hoboken, New Jersey 07030. Part of this work was carried out when the author had a visiting appointment at the Laboratory for Information and Decision Sciences at the Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

variable.[1] When interpreted as a spectral factorability type result, it allows us to embed an arbitrary positive or bounded (real) vector valued transfer function into a lossless positive or bounded (real) matrix. This is reminiscent of unitary dilation of contractive operators [14] and is well known in 1-D (infinite-dimensional) system theory. This key observation, along with synthesizability of lossless 2-D transfer function matrices [24]–[26], then demonstrate the synthesizability of arbitrary positive or bounded (real) transfer functions—a fact that has remained unavailable in the literature thus far. Since a passive synthesis automatically provides a dissipative realization, it in turn leads the way to a weak form of the 2-D KPY lemma and to our 2-D hyperstability results.

We wish to make a further observation on the 2-D aspect of our problem. It is well known that the local states such as those in 2-D Roesser's state space model [15] or in any other 2-D state model do not contain the full information regarding the complete history of the system. On the other hand, the global states are known to be infinite-dimensional. Although the details of our 2-D theory differ considerably from the theory of infinite-dimensional systems (which, in fact, are at least partially motivated by the study of partial differential/difference equations) analogous results, namely, those on linear quadratic optimal control [11], the KPY lemma [12], [13], including lossless embedding, i.e., unitary dilation of contractive operators [14], exist in this context. At a broad conceptual level these can, therefore, be taken to be supportive of the line of development reported in the present paper.

We now turn to applications in adaptive filtering. While recent literature has witnessed a profusion of activity in 1-D adaptive filtering both in the deterministic and stochastic context, in spite of its potential practical applicability, very little has been reported for higher dimensional problems. We refer to the books [34]–[38] and the recent review [33] for a survey of 1-D adaptive signal processing, whereas works of more control theoretic flavor are available, e.g., in [39], [41], [42]. Among these, the only deterministic recursive filtering scheme for which convergence of the adaptation algorithm can be theoretically demonstrated is known as the HARF (hyperstable adaptive recursive filter) and was first reported in [28]. Subsequently, the theory has been elaborated and further variations and simplifications are reported in [29]–[32], [36]. In this paper we show that the 2-D hyperstability theory developed in the first part of the paper can be crucially exploited in extending HARF, including its proof of convergence, to 2-D. Since we consider quarter plane causal filters, 2-D problems of the propagating wave type (specifically, those admitting hyperbolic differential/difference equation formulation), i.e., those arising in beamforming and/or target tracking, are expected to be particularly suitable problem areas where this theoretical study can find use. It may be mentioned that the need for adaptive filtering in image processing type 2-D problems has also been recently recognized (see, e.g., [44]–[46]) and the potential relevance of our work in this area, at least at a conceptual level, cannot be excluded by any means.

In §2 notation and terminologies are introduced first. In §3 a specific type of weak convergence of 2-D quarter plane signals relevant to our ensuing discussions on 2-D adaptive filtering is discussed. In §4 the hyperstability results are developed, while §5 introduces its applications to 2-D HARF type algorithms. In §6 discussions are presented and conclusions are drawn. An appendix is included to elaborate on Landau's theorem and its relevance to 2-D lossless embedding on which this entire

[1] The result also follows from combined use of Corollary 1.10 and the Cassel–Phister Theorem 1.3 in [9].

work is based.

2. Notation and terminology. Two-dimensional discrete signals are denoted by symbols such as $u(m,n)$ or $y(m,n)$ etc., where (m,n) are points in 2-D lattice space. All signals, unless otherwise specified or obvious from the context, are of quarter plane support. In the 2-D lattice space the following set of points will be of interest: the triangular set $T_N = \{(m,n);\ 0 \leq m+n \leq N\}$; the line set $C_N = \{(m,n);\ m+n = N\}$ and the trapezoidal set $R_{n_0 N} = \{(m,n);\ n_0 \leq m+n \leq N\} = T_N \setminus T_{n_0-1}$.

Z-transforms of 2-D signals, e.g., $u(m,n)$, $y(m,n)$ etc., will be denoted by $U(z_1, z_2)$, $Y(z_1, Z_2)$ etc., where z_1, z_2 are the transform variables. Similarly, p_1, p_2 will be used to denote the (Laplace) transform variables for continuous signals, which we will also feature in the present work. The (closed) unit bidisc of the complex biplane will be denoted by $(\bar{D})$ D, i.e., $\bar{D} = \{(z_1, z_2);\ |z_i| \leq 1, i = 1,2\}$ or $D = \{(z_1, z_2);\ |z_i| < 1, i = 1,2\}$ and its distinguished boundary $|z_1| = |z_2| = 1$ will be denoted by T. Also, the notation "Re" will denote the real part of a complex number. A real polynomial is one which has real valued coefficients, and a real rational function is the ratio of two real polynomials. The superscript $*$ denotes complex conjugation. If $g = g(z_1, z_2)$ is a rational function then its paraconjugate is $g_* = g^*(-p_1^*, -p_2^*)$ and if $a = a(z_1, z_2)$ is a rational function then its discrete paraconjugate is $\tilde{a} = a^*(z_1^{*^{-1}}, z_2^{*^{-1}})$. The partial degree of a two-variable polynomial g in the variable z_i is denoted by $\deg_i g$.

The notation $\|\cdot\|$ denotes a Eucledian norm of a row or a column vector. The norm $|H|$ of a matrix H denotes the spectral norm $|H| = \sup_{\|x\|} \|Hx\|$. A rational matrix (i.e., one with rational function entries) is real rational if its entries are so. The superscript T denotes Hermitian transpose, whereas superscript t denotes simply transpose of a matrix; $\tilde{}$ as a superscript or $*$ as a subscript to a rational matrix denotes the matrix with the corresponding operations performed on its entries and subsequently transposed (sometimes known as the parahermitian transpose). The notation $\oplus$ denotes the direct sum of matrices.

A rational function $g = g(p_1, p_2)$ is said to be *positive* if $\mathrm{Re} g \geq 0$, whenever $\mathrm{Re} p_i > 0$, $i = 1,2$. Similarly, $a = a(z_1, z_2)$ is *discrete positive* if $\mathrm{Re} a \geq 0$ in D. Also, $g = g(p_1, p_2)$ is *bounded* if $\|g\| \leq 1$, whenever $\mathrm{Re} p_i > 0$, i=1,2, and a is *discrete bounded* if $\|a\| \leq 1$ in D. If, in addition, the rational function concerned is a real rational function then it is included in the terminology by saying g is positive real, a is discrete positive real etc., as the case may be. Furthermore, a is said to *strictly discrete positive* (real) if it is discrete positive (real) and, in its irreducible form, neither its numerator nor its denominator has any zero in $\bar{D}$. Correspondingly, a is said to *strictly discrete bounded* (real) if it is discrete bounded (real) and $\|a\| < 1$ in $\bar{D}$.

3. Almost everywhere convergence of 2-D quarter plane double sequences. Our applications in adaptive 2-D IIR filtering of hyperstability theory require consideration of convergence of double sequences $x(m,n)$ in the quarter plane. We need to make precise a notion of convergence which does not demand that $x(m,n)$'s individually become small far away from the origin of the 2-D lattice space, but at "most" points they do in an areawise sense. We proceed as follows.

Let $\nu_\epsilon(C_N)$ be the number of points on C_N such that the 2-D signal $x(m,n)$ satisfies $|x(m,n)| > \epsilon$. Then $x(m,n)$ is said to be *almost everywhere convergent to zero* if for any n_0 and *arbitrary* ϵ we have

$$(1) \qquad \lim_{N\to\infty} \frac{\sum_{k=n_0}^{n_0+N} \nu_\epsilon(C_k)}{\ln(N+n_0+1)! - \ln n!} \le C < \infty,$$

where C is a constant independent of N.

The intuitive content of this definition is as follows. Consider the trapezoidal set R_{n_0N} in the 2-D lattice space, which, in fact, is of width $(N+1)$ with parallel sides of the trapezoid being the lines C_{n_0} and C_N. Note that the total number of points in R_{n_0N} is

$$(2) \qquad |R_{n_0N}| = \frac{1}{2}(N+1)(N+2n_0+2).$$

Before proceeding further we need to recall Stirling's formula [47] which estimates $n!$ for large n as follows:

$$\ln n! \approx (n-1)\ln n - n + \frac{1}{2}\ln n + \frac{1}{2}\ln 2\pi.$$

The essential point for us is that as $n \to \infty$, $\ln n!$ grows as $(n-1)\ln n$ which is faster than n but slower than n^2.

Thus as $N \to \infty$, the ratio

$$\begin{aligned} &\frac{1}{R_{n_0N}}\{\text{Number of points in } R_{n_0N} \text{ where } |x(m,n)| > \epsilon\} \\ = \quad & \frac{1}{R_{n_0N}} \sum_{k=n}^{N} \nu_\epsilon(C_k) \\ \le \quad & \frac{C}{R_{n_0N}}\{\ln(N+n_0+1)! - \ln n_0!\} \\ = \quad & C\frac{\ln(N+n_0+1)! - \ln n_0!}{\frac{1}{2}(N+1)(N+2n_0+2)} \\ \longrightarrow \quad & 0, \end{aligned}$$

where the first inequality follows from (1), the last equality follows from (2), and the last limit follows from the fact that, due to Stirling's formula, the numerator grows with N as $(N+n_0+1)\{\ln(N+n_0+1)-1\}$, whereas the denominator grows as N^2.

Thus, almost everywhere convergence to zero of $x(m,n)$ implies that the relative number of points in R_{n_0N}, where $|x(m,n)| > \epsilon$ with ϵ arbitrary, is small in the asymptotic limit $N \to \infty$.

Now let us define

$$(3) \qquad S_N = \sum\sum_{T_N} |x(m,n)|.$$

We then have the following result.

THEOREM 3.1. *If* (4) *holds for all integers* $N \ge 0$, *where* α, β *are constants independent of* N

$$S_N \leq N \cdot \alpha + \beta \tag{4}$$

then $x(m,n)$ is almost everywhere convergent to zero.

Proof. Fix $\epsilon > 0$ arbitrarily. Suppose for contradiction that $x(m,n)$ is not almost everywhere convergent to zero. Then for some fixed n_0 it is possible to find an integer N, arbitrarily large, such that

$$\sum_{k=n_0}^{n_0+N} \nu_\epsilon(C_k) > C\{\ln(N+n_0+1)! - \ln n_0!\} \tag{5}$$

for every fixed constant C. However, we also have from (3) and the definition of $\nu_\epsilon(C_k)$ that

$$S_{n_0+N} - S_{n_0-1} \geq \sum_{k=n_0}^{n_0+N} \epsilon\nu_\epsilon(C_k). \tag{6}$$

Combining (5) and (6) we have that for some fixed n_0 there exists arbitrarily large N such that the following inequality set holds true:

$$S_{n_0+N} - S_{n_0-1} > \epsilon C\{\ln(N+n_0+1)! - \ln n_0!\}, \tag{7}$$

i.e.,

$$S_{n_0+N} > S_{n_0-1} + \epsilon C \ln(N+n_0+1)! - \epsilon C \ln n_0!. \tag{8}$$

The second term in (8), by Stirling's formula, grows with N as $(N+n_0+1)\{\ln(N+n_0+1) - 1\}$, which is faster than the linear growth condition imposed on S_N by (4). Thus, (8) contradicts (4). Since $\epsilon > 0$ is arbitrarily small we have completed the proof of our result.

We next explore some useful properties of almost everywhere convergent sequences.

PROPOSITION 3.2. *Let $x(m,n)$ and $y(m,n)$ be both double sequences with quarter plane support each convergent to zero in the almost everywhere sense. Then $z(m,n) = x(m,n) + y(m,n)$ has the same property.*

Proof. Let $\nu_\epsilon(C_k)$ and $\mu_\epsilon(C_k)$ be the number of points on C_k where $|x(m,n)| > \epsilon$ and $|y(m,n)| > \epsilon$, respectively. Then if $\lambda_{2\epsilon}(C_k)$ denote the number of points on C_k, where $|z(m,n)| > 2\epsilon$ then we have

$$\lambda_{2\epsilon}(C_k) \leq \nu_\epsilon(C_k) + \mu_\epsilon(C_k).$$

Thus, we have

$$\begin{aligned}
&\lim_{N\to\infty} \frac{\sum_{k=n_0}^{n_0+N} \lambda_{2\epsilon}(C_k)}{\ln(N+n_0+1)! - \ln n_0!} \\
\leq\ & \lim_{N\to\infty} \frac{\sum_{k=n_0}^{n_0+N} \nu_\epsilon(C_k)}{\ln(N+n_0+1)! - \ln n_0!} + \lim_{N\to\infty} \frac{\sum_{k=n_0}^{n_0+N} \mu_\epsilon(C_k)}{\ln(N+n_0+1)! - \ln n_0!} \\
\leq\ & C_x + C_y < \infty,
\end{aligned}$$

where C_x and C_y are constants in (1) corresponding to almost everywhere convergence of $x(m,n)$ and $y(m,n)$, respectively. Since ϵ is arbitrary the almost everywhere convergence of $z(m,n)$ is demonstrated.

We will say that a double sequence is *asymptotically bounded almost everywhere* if for any finite n_0, (1) holds for *some* number ϵ. Clearly, a double sequence is asymptotically bounded almost everywhere if it converges to zero almost everywhere as $N \to \infty$.

We then have the following proposition.

PROPOSITION 3.3. *If $x(m,n)$ and $y(m,n)$ are two double sequences with quarter plane support such that $x(m,n)$ is almost everywhere convergent to zero and $y(m,n)$ is asymptotically bounded almost everywhere, then $z(m,n) = x(m,n)y(m,n)$ is almost everywhere convergent to zero.*

Proof. Let $\nu_\epsilon(C_k)$ be the number of points on C_k where $|x(m,n)| > \epsilon$ and $\mu_\kappa(C_k)$ denote the number of points C_k where $|y(m,n)| > \kappa$. Then we have

$$\lim_{N\to\infty} \frac{\sum_{k=n_0}^{n_0+N} \nu_\epsilon(C_k)}{\ln(N+n_0+1)! - \ln n_0!} \le C_x < \infty \tag{9}$$

$$\lim_{N\to\infty} \frac{\sum_{k=n_0}^{n_0+N} \mu_\kappa(C_k)}{\ln(N+n_0+1)! - \ln n_0!} \le C_y < \infty, \tag{10}$$

where (9) holds for arbitrary ϵ and (10) holds only for a *fixed* κ.

Now, let $\lambda_{\epsilon\kappa}(C_k)$ be the number of points on C_k where $|z(m,n)| > \epsilon\kappa$. Then $\lambda_{\epsilon\kappa}(C_k) \le \nu_\epsilon(C_k) + \mu_\kappa(C_k)$ and due to (9) and (10) we have

$$\lim_{N\to\infty} \frac{\sum_{k=n_0}^{n_0+N} \lambda_{\epsilon\kappa}(C_k)}{\ln(N+n_0+1)! - \ln n_0!} \le C_x + C_y < \infty. \tag{11}$$

Since ϵ is arbitrary and κ is fixed, the last equality proves that $z(m,n)$ is convergent to zero almost everywhere.

The following proposition follows almost immediately from the definition of almost everywhere boundedness.

PROPOSITION 3.4. *Let $x(m,n)$ and $y(m,n)$ be two double sequences with quarter plane support.*

1. *If $|x(m,n)| < K < \infty$, K =constant and $y(m,n)$ is asymptotically bounded almost everywhere then $z(m,n) = x(m,n) + y(m,n)$ is also asymptotically bounded almost everywhere.*
2. *If $x(m,n)$ and $y(m,n)$ are both asymptotically bounded almost everywhere then so is $z(m,n) = x(m,n)y(m,n)$.*

4. 2-D hyperstability type results. In this section we develop the main results on 2-D extensions of hyperstability theory. Although many other variations of the results could be presented and further generalizations of them are possible, our main goal is to arrive at statements of certain facts that would allow us to prove the convergence of the adaptive algoritm to be discussed in §5.

Before proceeding further it is important to emphasize an important distinction between 1-D and 2-D systems. Among the many possible ways in which the "successive" states of a 2-D quarter plane system can be updated from "previous" states, we focus on the one in which all states on C_N are simultaneously computed from those on C_{N-1}. For computation of C_N from C_{N-1}, boundary values of the state variables, namely those at $(0, N)$ and at $(N, 0)$, need to be specified. Note that these do not influence the evolution of the system until states on C_N have been computed. Thus, unlike in 1-D, *new* boundary values of states can keep adding information to the system not previously used and can thus influence the output as computation progresses. Consequently, notions related to stability, e.g., boundedness, etc., of output and/or state may, in general, require qualifications as to the specified boundary values of the states. However, in most of what follows we validly assume, unless otherwise specified, that the boundary values of states are zero and thus, the evolution of the system is completely governed by its zero state response, i.e., by its transfer function only.

The main results of this section are the 2-D hyperstability theorems stated in Theorems 4.1 and 4.2.

THEOREM 4.1 (Bounded version). *Let $H = H(z_1, z_2)$ be the scalar transfer function of a 2-D quarter plane causal system. Let H be strictly discrete bounded (real) (i.e., $|H| < 1$ in $\bar{D}$), $u(m,n)$ be an input, $y(m,n)$ be the corresponding output from the system while the boundary conditions are assumed zero. Also, let*

$$S_N = \sum_{T_N}\sum |u(m,n)|^2 - |y(m,n)|^2. \tag{12}$$

Then we have the following:

(i) *If $S_N \le K_1 N + K_2$, where K_1 and K_2 are constants independent of N then for some β with $K_1' = |\beta|^2 K_1$ and $K_2' = |\beta|^2 K_2$ we have*

$$\sum_{T_N}\sum |u(m,n)|^2 \le K_1' N + K_2'; \ \sum_{T_N}\sum |y(m,n)|^2 \le K_1' N + K_2'. \tag{13}$$

Consequently, both $y(m,n)$ and $u(m,n)$ converge to zero almost everywhere.

(ii) *If $S_N \le K < \infty$ for every nonnegative integer N then we have that $y(m,n) \to 0$ and $u(m,n) \to 0$ as $(m+n) \to \infty$ in the usual sense.*

Before undertaking a detailed proof of Theorem 4.1 we will make several comments on its intuitive contents and on our strategy for its proof. First, note that a realization of a (strictly) discrete bounded transfer function can be viewed as a (strictly) dissipative system, whereas the expression for S_N clearly indicates that it stands for the total flow of energy into the system minus the total flow of energy out of the system over the triangular region T_N. Thus the boundedness condition $S_N \le K < \infty$ implies that the net flow of energy into the system from outside is at most K. Now, since the system is strictly dissipative, if there is no energy input from the boundary values of the state variables (i.e., if they are assumed to be zero), in order for the energy conservation law to hold, the energy stored in the state variables must die out to zero; consequently, the output must also die out to zero. This last arguement, in fact, requires minimality of the realization so that self-sustaining oscillatory modes, which are neither uncontrollable from the input nor unobservable from the output, are excluded. However, the conclusion of the theorem remains true even if the realization is nonminimal. This latter point is vital in our proof of Theorem 4.1 because of the fact that in 2-D, minimal realization, its characterization, and its properties are not yet clearly understood [19].

Second, if $S_N \leq K_1 N + K_2$, i.e., if the net flow of energy into the system is not bounded but can only grow at most linearly with (the propagation of the "computational wavefront") N, the asymptotic decay of the output cannot be argued as above, and in fact, is not even true. However, we do have convergence of $y(m,n)$ to zero in the almost everywhere sense as discussed in §3. Here we may only roughly mention that since the "area" of the triangle T_N increases as N^2 with N, a linear rate of inflow of energy cannot keep up with the increase in area so that it is, in an average sense, thinly distributed over the 2-D lattice space for large N.

Third, although we have assumed that the boundary values of the states, i.e., those at points $(m,0)$ and $(0,n)$ for nonnegative m and n are zero, it is not exactly necessary for the validity of Theorem 4.1 (the same holds for Theorem 4.2 to follow). It should be plausible for the above discussions and would be more transparent in the proof that both conclusions (i) and (ii) of Theorem 4.1 remain valid if we assume that the boundary values of the states are square summable, i.e., only a finite amount of energy is fed into the system from boundary values of states, while (i) only remains valid if we allow this energy to grow at most linearly with N.

A proof of the above theorem will be given essentially via circuit theoretic arguments. In particular, we draw from existing results on structurally passive synthesis of 2-D lossless trnasfer functions [24]–[26] and a little known 2-D spectral factorizability type result by Landau [7]. Landua's result allows us to embed an arbitrary 2-D bounded (real) or positive (real) transfer function into a 2-D lossless scattering or immittance matrix. The embedding, in physical terms, can be viewed as the operation of resistor extraction [5], thus leaving a lossless multiport to be synthesized via procedures estabilished in [24]–[26]. This then yields a structurally passive synthesis of arbitrary 2-D bounded (real) transfer functions. Note that in 1-D, a minimal passive synthesis along with the state space isomorphism result [2] directly yields the positive real (KPY) lemma, or the bounded real lemma, which in turn can be crucially exploited to provide proofs of the 1-D hyperstability theorems as elaborated in [2] and [4]. Although in our 2-D context we do not necessarily have a minimal passive synthesis via the procedure outlined above, and a state space isomorphism result is not known, a weak 2-D version of the hyperstability theorem useful in the present context of adaptive 2-D signal processing, namely that stated in Theorems 4.1 and 4.2, can indeed be proved. Theorem 4.1 and its counterpart in [28] are weak in the sense that we only require the input $u(m,n)$ and the output $y(m,n)$, but not necessarily the states in a realization of H to be asymptotically zero, while the original formulation of the hyperstability theorem in [1], [2], and [4] requires, in addition, that the state variables in a minimal realization of H necessarily converge to zero.

To proceed we first establish the following result.

LEMMA 4.2. *Let $H = H(z_1, z_2)$ be a rational matrix of size $(1 \times \ell)$, which, in addition, is discrete bounded. Then there exist integers p, q and matrices A, B, C, and D of appropriate sizes such that*

$$H(z_1, z_2) = A + B(z_1^{-1} I_p \oplus z_2^{-1} I_q - D)^{-1} C \tag{14}$$

along with

$$I - T^* T \geq 0, \tag{15}$$

where

$$T = \begin{bmatrix} A & B \\ C & D \end{bmatrix}. \tag{16}$$

Furthermore, if H is real rational then T is real.

Remark 4.1. Note that Lemma 4.2 essentially states that a *passive* 2-D Roesser's state space realization for the transfer function $H(z_1, z_2)$ can be obtained. If B, C, D are partitioned as

$$D = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix}; B = \begin{bmatrix} B_1 & B_2 \end{bmatrix}; C = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix}, \tag{17}$$

where D_{11}, D_{12} are of respective sizes $(p \times p)$; B_1, B_2 has, respectively, p and q columns; C_1, C_2 has, respectively, p and q rows, then $H(z_1, z_2)$ can be viewed as the transfer function between the input vector $u(m,n)$ and the output $y(m,n)$ in the Roesser's state space model:

$$\begin{bmatrix} x_h(m+1,n) \\ x_v(m,n+1) \end{bmatrix} = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} x_h(m,n) \\ x_v(m,n) \end{bmatrix} + \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} u(m,n), \tag{18}$$

$$y(m,n) = \begin{bmatrix} B_1 & B_2 \end{bmatrix} \begin{bmatrix} x_h(m,n) \\ x_v(m,n) \end{bmatrix} + Au(m,n). \tag{19}$$

A realization satisfying (15) to (19) will henceforth be called a "*passive realization*" of a discrete bounded H.

Remark 4.2. The matrix inequality in (15) is equivalent to the existence of two further matrices L and W such that

$$1 - T^*T = \begin{bmatrix} L^* \\ W^* \end{bmatrix} \begin{bmatrix} L & W \end{bmatrix}, \tag{20}$$

where L is of size $(r \times \ell)$ and W is of size $(r \times (p+q))$ for some $r \leq p+q+1$. Equations (15) and (16) together can be viewed as a weak form of a 2-D bounded real lemma for the transfer function H. It is weak in that it applies only to the specific realization obtained in Lemma 4.2. The lack of a 2-D state space isomorphism result prevents further generalization to arbitrary "minimal" realizations in 2-D. Note further that (15) and (16) combined together can be written in a more familiar form:

$$I_p - A^*A - C^*C = L^*L, \tag{21}$$

$$-A^*B - C^*D = L^*W, \tag{22}$$

$$I_q - B^*B - D^*D = W^*W. \tag{23}$$

COROLLARY 4.3. *Let $H = H(z_1, z_2)$ be a (real) rational matrix of size $(1 \times \ell)$, which, in addition, is discrete bounded. Let A, B, C, D as in (16) be a passive* *realization of H. Let L and W be defined as in* (20) *or, equivalently, in* (21)–(23). *Then we have*

$$1 - \tilde{H}(z_1, z_2)H(z_1, z_2) = \tilde{P}(z_1, z_2)P(z_1, z_2), \tag{24}$$

where $P(z_1, z_2)$ *is the transfer function of the Roesser's state space model given by*

$$\begin{bmatrix} x_h(m+1, n) \\ x_v(m, n+1) \end{bmatrix} = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \begin{bmatrix} x_h(m, n) \\ x_v(m, n) \end{bmatrix} + \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} u(m, n), \tag{25}$$

$$\eta(m, n) = \begin{bmatrix} W_1 & W_2 \end{bmatrix} \begin{bmatrix} x_h(m, n) \\ x_v(m, n) \end{bmatrix} + Lu(m, n) \tag{26}$$

with $W = \begin{bmatrix} W_1 & W_2 \end{bmatrix}$. *In other words,* $P(z_1, z_2)$ *is given by*

$$P(z_1, z_2) = L + W[z_1^{-1} I_p \oplus z_2^{-1} I_q - D]^{-1} C. \tag{27}$$

Proof. The proof is routine algebraic manipulation for which we adopt the compact notation $\zeta I = z_1 I_p \oplus z_2 I_q$. By substituting for H from (14) and expanding the product, then from (21),(22),(23) by making the replacements $I_p - A^*A = C^*C + L^*L$, $A^*B = -(C^*D + L^*W)$, $B^*B = I_q - D^*D - W^*W$ and subsequently rearranging terms we can write

$$1 - \tilde{H}(z_1, z_2)H(z_1, z_2) = \tilde{P}(z_1, z_2)P(z_1, z_2) + C^*(t_1 + t_2 + t_3)C, \tag{28}$$

where

$$t_1 = I - (\zeta I - D^*)^{-1}(\zeta^{-1} I - D)^{-1},$$

$$t_2 = D(\zeta^{-1} I - D)^{-1} + (\zeta I - D^*)^{-1} D^*,$$

and

$$t_3 = (\zeta I - D^*)^{-1} D^* D (\zeta^{-1} I - D)^{-1}.$$

Since we obviously have

$$\begin{aligned} t_1 &= (\zeta I - D^*)^{-1}\{(\zeta I - D^*)(\zeta^{-1} I - D) - I\}(\zeta^{-1} I - D)^{-1} \\ &= (\zeta I - D^*)^{-1}\{D^* D - D^* \zeta^{-1} - D\zeta\}(\zeta^{-1} I - D)^{-1} \end{aligned}$$

and

$$\begin{aligned} t_2 &= (\zeta I - D^*)^{-1}\{D^*(\zeta^{-1} I - D) + (\zeta I - D^*)D\}(\zeta^{-1} I - D)^{-1} \\ &= (\zeta I - D^*)^{-1}\{D^* \zeta^{-1} + D\zeta - 2D^* D\}(\zeta^{-1} I - D)^{-1} \end{aligned}$$

it follows after further expansion that $t_1 + t_2 + t_3 = 0$, thus proving the corollary.

Proof of Lemma 4.2. It follows from the 2-D discrete embedding Theorem (cf. Theorem A.4) that there exists a 2-D discrete lossless bounded (real) matrix $G(z_1, z_2)$ of size $(s+\ell) \times (s+\ell)$ such that the top left $(1 \times \ell)$ block of G is $H(z_1, z_2)$. We thus have the situation depicted in Fig. 1, in which

$$\begin{bmatrix} Y(z_1,z_2) \\ Y'(z_1,z_2) \end{bmatrix} = G(z_1,z_2) \begin{bmatrix} U(z_1,z_2) \\ U'(z_1,z_2) \end{bmatrix} \tag{29}$$

$$G = G(z_1,z_2) = \begin{bmatrix} H & G_{12} \\ G_{21} & G_{22} \end{bmatrix}, \tag{30}$$

where G_{22} is of size $(\ell + s - 1) \times s$ and capital letters are used to denote transform domain variables corresponding to small english letters.

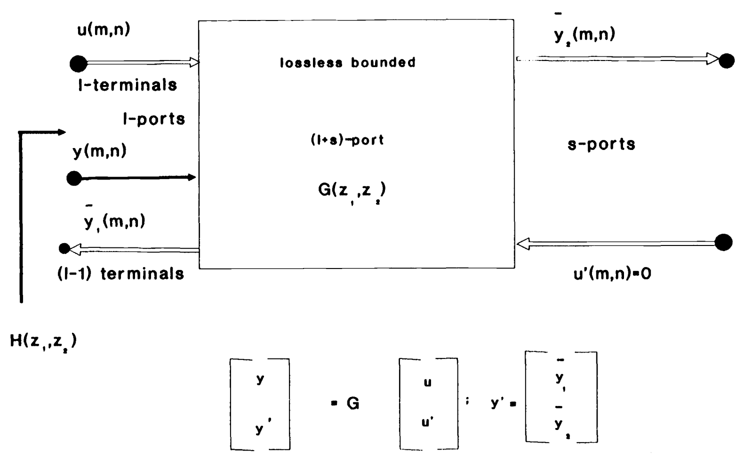

u(m,n)
l-terminals
l-ports
y(m,n)
$\bar{y}_1(m,n)$
(l-1) terminals
H(z₁,z₂)
lossless bounded
(l+s)-port
G(z₁,z₂)
$\bar{y}_2(m,n)$
s-ports
u'(m,n)=0
$\begin{bmatrix} y \\ y' \end{bmatrix} = G \begin{bmatrix} u \\ u' \end{bmatrix}$; $y' = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \end{bmatrix}$

FIG. 1

Now, it is known that [24]–[26] it is possible to obtain a passive (in fact, minimal) synthesis for any 2-D discrete lossless bounded matrix, and thus of $G(z_1, z_2)$ in particular. More specifically, one way of viewing the synthesis procedure is to extract p of z_1-type delays and q of z_2 delays in such a way that we are left with a constant lossless bounded multiport with as many as $(p+q+s+\ell)$ ports. Let the transfer function of the constant lossless bounded $(p+q+s+\ell)$-port be given by (cf. Fig. 2) $\bar{T}$ as

$$\begin{bmatrix} y(m,n) \\ x_h(m+1,n) \\ x_v(m,n+1) \\ y'(m,n) \end{bmatrix} = \bar{T} \begin{bmatrix} u(m,n) \\ x_h(m+1,n) \\ x_v(m,n+1) \\ u'(m,n) \end{bmatrix} \tag{31}$$

and partition $\bar{T}$ as

$$\bar{T} = \begin{bmatrix} A & B & T_{13} \\ C & D & T_{23} \\ T_{31} & T_{32} & T_{33} \end{bmatrix}, \tag{32}$$

where A is $(1 \times \ell)$, D is $(p+q)$ square, T_{33} is $(\ell + s - 1) \times s$. Since $\bar{T}$ is lossless bounded we have $\bar{T}^*\bar{T} = I$, which in view of (32) yields (33) in which T is as in (16)

$$I - T^*T \geq 0. \tag{33}$$

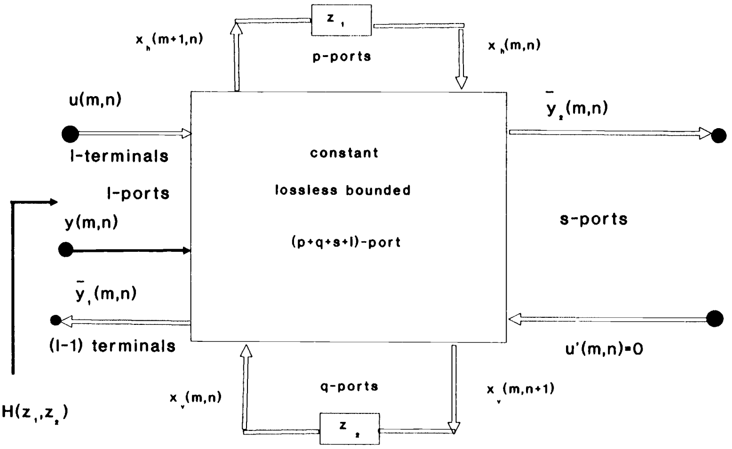


FIG. 2

Clearly, since in Fig. 2 $x_h(m,n)$ and $x_v(m,n)$ are outputs of z_1 and z_2 type delays, respectively, they are valid state variables. Furthermore, by setting $u'(m,n) = 0$ it can be verified from (32) that T obtained as above then indeed corresponds to Roesser's state space model (18) and (19). It then is well known that the transfer function H between $u(m,n)$ and $y(m,n)$ is indeed given by the formula (14), thus completing the proof of the present lemma.

We will need the following notions for the developments to follow.

Remark 4.3. Let $H = H(z_1, z_2)$ be a (real) rational matrix of size $(1 \times \ell)$, holomorphic in $\bar{D}$. If $\alpha = sup_{\bar{D}}|H(z_1, z_2)|$ then clearly α is a well-defined finite real number. We then consider $H_\alpha = H_\alpha(z_1, z_2) = \alpha^{-1}H(z_1, z_2)$. Obviously, then $\|H_\alpha\| \leq 1$ in $\bar{D}$. Thus, H_α is a discrete bounded (real) matrix for which there exists a passive realization as in Remark 4.1. This, in fact, also gives us a realization of H, if the A-matrix and the C-matrix in (18) and (19) are replaced by αA and αC, respectively. Such a realization of a transfer function H, holomorphic in $\bar{D}$, will also be referred to as a "*passive realization.*"

We can then state the following result.

LEMMA 4.4. *Let $H = H(z_1, z_2)$ be a rational discrete bounded (real) transfer function of a passive realization having ℓ-inputs and 1-output operating under zero boundary conditions. Let $u(m,n)$ be the input vector and $y(m,n)$ be the corresponding output, whereas let $\eta(m,n)$ be given by*

$$\eta(m,n) = W\phi(m,n) + Lu(m,n), \tag{34}$$

where

$$\phi(m,n) = \begin{bmatrix} x_h(m,n) \\ x_v(m,n) \end{bmatrix}$$

is the state vector in the passive realization of H. We then have the following:

(i) *If for all nonnegative integer N*

$$\sum_{T_N}\sum \|u(m,n)\|^2 \leq K_1N + K_2 \tag{35}$$

then for all nonnegative integer N

$$\sum_{T_N}\sum |y(m,n)|^2 \leq K_1N + K_2, \tag{36}$$

where K_1 and K_2 are constants independent of N.

(ii) *If we have that for all nonegative integer N*

$$S_N = \sum_{T_N}\sum [\|u(m,n)\|^2 - |y(m,n)|^2] \leq K_1N + K_2 \tag{37}$$

then necessarily for all nonnegative integer N we have

$$\sum_{T_N}\sum \|\eta(m,n)\|^2 \leq K_1N + K_2. \tag{38}$$

Proof. Consider a passive realization of H as in (18),(19). If in this realization we denote by $\hat{\phi}(m,n)$ the vector

$$\hat{\phi}(m,n) = \begin{bmatrix} x_h(m+1,n) \\ x_v(m,n+1) \end{bmatrix}, \tag{39}$$

then we have from (18)

$$\begin{aligned} \|\hat{\phi}(m,n)\|^2 &= [D\phi(m,n) + Cu(m,n)]^*[D\phi(m,n) + Cu(m,n)] \\ &= \|\phi(m,n)\|^2 - \|\eta(m,n)\|^2 + \|u(m,n)\|^2 - |y(m,n)|^2, \end{aligned} \tag{40}$$

where the last step follows via the use of (22),(23),(19), and (26). Next, by summing (40) over the set of all (m,n) in T_N we have

$$
\begin{aligned}
\sum_{T_N}\sum \|\hat{\phi}(m,n)\|^2 &= \sum_{T_N}\sum \|\phi(m,n)\|^2 - \sum_{T_N}\sum \|\eta(m,n)\|^2 \\
&+ \sum_{T_N}\sum [\|u(m,n)\|^2 - |y(m,n)|^2]. \qquad (41)
\end{aligned}
$$

Recalling the definition of $\hat{\phi}(m,n)$ and $\phi(m,n)$ in terms of $x_h(m,n)$, $x_v(m,n)$ and then cancelling the identical terms in the left- and right-hand sides of (41) and then finally by adding the terms $x_v(0,N)$ and $x_h(N,0)$ to both sides, we can write

$$
\begin{aligned}
\sum\sum\nolimits_{C_N} \|\phi(m,n)\|^2 &= \sum_{m=0}^{N} \|x_h(m,0)\|^2 + \sum_{n=0}^{N} \|x_v(0,n)\|^2 \\
&- \sum_{T_N}\sum \|\eta(m,n)\|^2 \\
&+ \sum_{T_N}\sum [\|u(m,n)\|^2 - |y(m,n)|^2]. \qquad (42)
\end{aligned}
$$

Note that the above is a manifestation of energy conservation law. The first term in the left-hand side is a measure of energy stored in the realization at the Nth stage of computation, the first two terms in the right-hand side are measures of energy coming from boundary conditions, whereas the second and the third terms, respectively, represent the total dissipation out of and the net flow of energy into the system over the entire region T_N of the 2-D lattice space.

Since the boundary conditions are assumed zero, i.e., $x_h(m,0) = x_v(0,n) = 0$ for all m, n we have the following obvious inequality set:

$$
\sum_{T_N}\sum |y(m,n)|^2 \le \sum_{T_N}\sum \|u(m,n)\|^2 \qquad (43)
$$

and

$$
\sum_{T_N}\sum \|\eta(m,n)\|^2 \le \sum_{T_N}\sum [\|u(m,n)\|^2 - |y(m,n)|^2]. \qquad (44)
$$

Conclusion (i) of the present lemma then follows from (43), whereas conclusion (ii) follows from (37) and (44).

Remark 4.4. If, instead of $x_h(m,0) = x_v(0,n) = 0$ as in the above theorem, we have

$$
\sum_{m=0}^{\infty} \|x_h(m,0)\|^2 \le C_h < \infty; \ \sum_{n=0}^{\infty} \|x_v(0,n)\|^2 \le C_v < \infty, \qquad (45)
$$

then it is clear from (42) that inequalities (43) and (44) still remain valid with an additional finite term $(C_h + C_v)$ in their right-hand sides. The conclusion of Lemma 4.4 thus still holds under this less restrictive situation. The physical implication of this is that if the total energy fed by the boundary values to the system is finite then the hyperstability theorem remains true.

For a complete proof of Theorem 4.1 we will need the following sequence of results of somewhat technical nature.

LEMMA 4.5. *Let $H = H(z_1, z_2)$ be a rational matrix of size $(\ell \times 1)$ which is holomorphic as well as nonzero in $\bar{D}$. Then there exists a rational matrix $G = G(z_1, z_2)$ of size $(1 \times \ell)$ such that $GH = 1$ and each entry of G is holomorphic in $\bar{D}$, i.e., H has a "stable" left inverse G. Furthermore, if H is real rational then so is G.*

Proof. We can write $H = [\phi_{i1}]/d$, where ϕ_{i1}, for $i = 1$ to ℓ and d are polynomials. Clearly, then $d \neq 0$ and the ϕ_{i1}'s are not simultaneously zero in $\bar{D}$.

Next, let (ζ_i, ω_i); $i = 1$ to m (we assume $m < \infty$; otherwise, due to Bezout's theorem in algebraic geometry, the ϕ_{i1}'s must have a common factor, after the extraction of which the essential problem can once again be reduced to the case $m < \infty$) be the common zeros of the set of polynomials ϕ_{i1}; $i = 1$ to ℓ. Let us order these zeros such that $|\zeta_i| > 1$ for $i = 1$ to ν and $|\omega_i| > 1$ for $i = \nu + 1$ to m and then consider the real polynomial

$$\pi = \pi(z_1, z_2) = \prod_{i=1}^{\nu}(z_1 - \zeta_i)(z_1 - \zeta_i^*) \prod_{i=\nu+1}^{m} (z_2 - \omega_i)(z_2 - \omega_i^*).$$

Clearly, π is zero whenever the ϕ_i's simultaneously vanish. Thus, by invoking Hilbert's Nullstellensatz [10, p. 5] it follows that there exists a nonnegative integer μ such that π^μ belongs to the ideal generated by ϕ_{i1}; $i = 1$ to ℓ. That is, there exist polynomials ψ_{1i}; $i = 1$ to ℓ such that $\sum_{i=1}^{\ell} \phi_{i1}\psi_{1i} = \pi^\mu$. Since π is clearly nonzero in $\bar{D}$ the desired result immediately follows by defining $G = [G_{1i}]$ with $G_{1i} = \psi_{1i}/\pi^\mu$ for $i = 1$ to ℓ. Furthermore, if H is real rational then it also follows from Nullstellensatz that ψ_{1i}'s are real polynomials and then G_{1i}'s are real rational functions.

LEMMA 4.6. *If $G = G(z_1, z_2)$ is the (real) rational transfer function of a one-input ℓ-output system and G is holomorphic in $\bar{D}$ then there exists a constant α such that $G_1 = \frac{1}{\alpha}G$ is discrete bounded (real), and thus, G admits a passive realization.*

Proof. Since G is holomorphic in $\bar{D}$, $\alpha = \sup_{\bar{D}} |G| < \infty$. Define then $G_1 = \frac{1}{\alpha}G$, which obviously satisfies $\sup_{\bar{D}} |G_1| = 1$. In particular, $\sup_T |G_1| \leq 1$ i.e., $(I_\ell - \tilde{G}_1 G_1) \geq 0$ on T. This, along with holomorphy of G_1 in $\bar{D}$ imply (cf. Theorem 28 in [23]) that G_1 is a discrete bounded (real) matrix. Thus, in view of Remark 4.1, G_1 admits a passive realization, which, due to $G = \alpha G_1$ and Remark 4.3, implies that G also admits a passive realization.

COROLLARY 4.7. *If H is a (real) rational matrix of size $(1 \times \ell)$ such that it is holomorphic as well as nonzero in $\bar{D}$ then there is a (real) rational left inverse G of H that admits a passive realization.*

Proof. Choose G to be the left inverse of H as prescribed in Lemma 4.5. Since such a G is neccesarily holomorphic in $\bar{D}$, due to Lemma 4.6, it admits a passive realization.

LEMMA 4.8. *Let $G = G(z_1, z_2)$ be a $(1 \times \ell)$ (real) rational matrix which is holomorphic in $\bar{D}$, $u(m, n)$ be the input vector, $y(m, n)$ be the corresponding output from a passive realization of $G(z_1, z_2)$ operating under zero boundary conditions. Then a linear bound on the input energy*

$$\sum_{T_N}\sum \|u(m,n)\|^2 \leq K_1 N + K_2 \quad \text{for all } N \tag{46}$$

implies the following linear bound on the output energy:

(47) $$\sum_{T_N}\sum |y(m,n)|^2 \le (K_1N + K_2)|\alpha|^2 \quad \textit{for all } N,$$

where $\alpha = sup_{\bar{D}}|G(z_1, z_2)| < \infty$ *and* K_1 *and* K_2 *are constants independent of* N.

Proof. Note that in view of Remark 4.3 under the conditions stated $G_\alpha(z_1, z_2) = \frac{1}{\alpha}G(z_1, z_2)$ is a discrete bounded (real) transfer function and a passive realization of G_α is obtained by replacing the A-matrix and the C-matrix in that for G by $A' = \alpha A$ and $C' = \alpha C$, respectively. Equivalently, the two realizations can be viewed as identical with the input for G_α being replaced by $u_\alpha(m,n) = \alpha u(m,n)$. We thus have from (46) that

(48) $$\sum_{T_N}\sum \|u_\alpha(m,n)\|^2 \le (K_1N + K_2)|\alpha|^2.$$

Since G is discrete bounded (real) it follows by invoking Lemma 4.4(ii) that the output $y(m,n)$ from G due to the input $u(m,n)$, which is the same as the output from G_α due to the input $u_\alpha(m,n)$ satisfies the inequality (47).

LEMMA 4.9. *Let* $H = H(z_1, z_2)$ *be a* $(\ell\times 1)$ *rational discrete bounded (real) matrix holomorphic as well as nonzero in* $\bar{D}$. *If the output vector* $y(m,n)$ *from a one-input* ℓ*-output system with transfer function* H *operating under zero boundary conditions corresponding to an input* $u(m,n)$ *is linearly bounded as*

(49) $$\sum_{T_N}\sum \|y(m,n)\|^2 \le K_1N + K_2,$$

then the input $u(m,n)$ *must also be linearly bounded as*

(50) $$\sum_{T_N}\sum |u(m,n)|^2 \le K_1'N + K_2',$$

where $K_1' = |\beta|^2K_1$, $K_2' = |\beta|^2K_2$ *and* K_1, K_2, β *are constants independent of* N.

Proof. Consider a left inverse G of H such that G admits a passive realization. Such a G exists due to Corollary 4.7. Since G is left inverse of H, $u(m,n)$ can be viewed as the output from this passive realization of G, when driven with input $y(m,n)$ under zero boundary conditions. In fact, as in the proof of Lemma 4.8 we have $G_1 = \frac{1}{\beta}G$, where $\beta = \sup_{\bar{D}}|G| < \infty$, G_1 is discrete bounded (real), and realization of G is obtained by multiplying the A-matrix and the C-matrix in that of G_1 by β. Thus, from Lemma 4.4(i) it follows that the output $u(m,n)$ obeys the linear bound (50) with $K_1' = |\beta|^2K_1$, $K_2' = |\beta|^2K_2$.

We now turn to the proof of Theorem 4.1.

Proof of Theorem 4.1. Consider a passive realization of $H(z_1, z_2)$ as in (18) and (19) with associated $\eta(m,n)$ as in (26), where L and W are defined via (20) or (21)–(23). We then have from part (ii) of Lemma 4.4 that if $S_N \le K_1N + K_2$ then

(51) $$\sum_{T_N}\sum \|\eta(m,n)\|^2 \le K_1N + K_2.$$

Next, note that $\eta(m,n)$ can be viewed, as in Corollary 4.3, as the output from a Roesser's state space model described by (25) and (26), which has the transfer function $P(z_1,z_2)$ as in (27). Since $H(z_1,z_2)$ is assumed to be strictly bounded, i.e., $|H| < 1$ in $\bar{D}$, we conclude from (24) that $\tilde{P}P \neq 0$ in $\bar{D}$. In particular, $P \neq 0$ in $\bar{D}$. Also, since H is strictly bounded (real), P along with H is holomorphic in $\bar{D}$. Thus, by likening P with H, and $\eta(m,n)$ with $y(m,n)$ it follows, in view of (51), from Lemma 4.9 that the input $u(m,n)$ to $P(z_1,z_2)$ is linearly bounded as

$$\sum_{T_N}\sum |u(m,n)|^2 \leq K_1'N + K_2', \tag{52}$$

where $K_1'/K_1 = K_2'/K_2 = |\beta|^2$ is a constant as in Lemma 4.9.

Furthermore, since $u(m,n)$ is the input and $y(m,n)$ is the output from the realization (18) and (19) (operating under zero initial conditions), whose transfer function $H = H(z_1,z_2)$ is strictly discrete bounded (real) it follows by again applying Lemma 4.4 (i) that

$$\sum_{T_N}\sum |y(m,n)|^2 \leq K_1'N + K_2'. \tag{53}$$

Part (ii) of Theorem 4.1 corresponds to assuming $K_1 = 0$, $K_2 = K$. Thus, in this case it follows from (53) that $y(m,n) \to 0$ as $(m+n) \to \infty$ in the usual sense. Furthermore, if $K_1 \neq 0$ then the result of part (i) follows by invoking Theorem 3.1 on almost everywhere convergence along with (53).

Alternative strategy of proof of Theorem 4.1. Consider $H(z_1,z_2) = \sum_{r=0}^{\infty} h_r(z_1,z_2)$, where h_r is a homogeneous polynomial of degree r. If H is strictly bounded real rational then $s(\lambda) = H(\lambda z_1, \lambda z_2)$ is a bounded function of λ for fixed z_1, z_2 in the closed unit bidisc.

If $Y(z_1,z_2) = \sum_{m,n}^{\infty} y(m,n)z_1^m z_2^n = \sum_{r=0}^{\infty} y_r(z_1,z_2)$ is the output corresponding to the input $U(z_1,z_2) = \sum_{m,n}^{\infty} u(m,n)z_1^m z_2^n = \sum_{r=0}^{\infty} u_r(z_1,z_2)$, where y_r and u_r are homogeneous polynomials of degree r, then clearly $y_0+\lambda y_1+\cdots = (h_0+\lambda h_1+\cdots)(u_0+\lambda u_1+\cdots)$. By equating coefficients of similar powers of λ we obtain $\tilde{y}_N = \Delta_N \cdot \tilde{u}_N$, where $\tilde{u}_N = [u_0, u_1, \cdots u_N]^t$, $\tilde{y}_N$ is defined analogously, and Δ_N is the lower triangular Toeplitz matrix with its first column equal to $[h_0, h_1, \cdots h_N]^t$. Thus, for any N,

$$||\tilde{u}_N||^2 - ||\tilde{y}_N||^2 = \tilde{u}_N^*(I_{N+1} - \Delta_N^*\Delta_N)\tilde{u}_N \geq (1-\mu_N)||\tilde{u}_N||^2 > 0, \tag{54}$$

where the norm $||\cdot||$ is computed for $z_1 = z_2 = \exp(j\omega)$; ω =real, and μ_N is the largest eigenvalue of $\Delta_N^*\Delta_N$. The last inequality in (54) is a consequence of boundedness of $s(\lambda)$ [16], from which it also follows that there exists an α such that $\mu_N < \alpha^2 < \infty$ for all real ω. Thus,

$$||\tilde{u}_N||^2 \leq \frac{||\tilde{u}_N||^2 - ||\tilde{y}_N||^2}{1-\alpha^2}. \tag{55}$$

But since

$$\left(\frac{1}{2\pi}\right)^2 \int_0^{2\pi}\int_0^{2\pi} ||\tilde{u}_N(e^{j\omega_1}, e^{j\omega_2})||^2 d\omega_1 d\omega_2 = \sum_{T_N}\sum |u(m,n)|^2, \tag{56}$$

it follows that boundedness of S_N as in (12) by $K_1N + K_2$ implies the first equation (13) with $K_i' = K_i/(1 - \alpha^2)$ for $i = 1, 2$, whereas the second equation (13) follows from (54).

For the purpose of application of the 2-D hyperstability theorem to adaptive signal processing in the next section a version different from that in Theorem 4.1 proves to be more appropriate.

THEOREM 4.10 (Positive version). *Let $F = F(z_1, z_2)$ be the scalar rational transfer function of a 2-D quarter plane causal system. Let F be strictly discrete positive (real) (i.e., $\mathrm{Re}F > 0$ in D), $p(m,n)$ be an input , $q(m,n)$ be the corresponding output from the system while the boundary conditions are assumed to be zero . Also let*

$$E_N = \mathrm{Re}\left\{\sum_{T_N}\sum p^*(m,n)q(m,n)\right\}. \tag{57}$$

Then we have the following:

(i) *If $E_N \leq K - 1N + K_2$, where K_1 and K_2 are constants independent of N then*

$$\sum_{T_N}\sum |p(m,n)|^2 \leq K_1'N + K_2',$$

$$\sum_{T_N}\sum |q(m,n)|^2 \leq K_1'N + K_2',$$

where $K_1' = |\beta|^2K_1, K_2' = |\beta|^2K_2$, and β are finite constants independent of N. Consequently, $p(m,n)$ and $q(m,n)$ converge to zero almost everywhere in the sense decribed earlier.

(ii) *If $E_N \leq K < \infty$ for every nonnegative integer N then we have that $p(m,n) \to 0$ and $q(m,n) \to 0$ as $(m+n) \to \infty$ in the usual sense.*

Proof. Consider signals $u(m,n) = p(m,n) + q(m,n)$ and $y(m,n) = P(m,n) - q(m,n)$, which can be viewed, respectively, as the input and the corresponding output from a system with transfer function $H = (1 - F)/(1 + F)$ operating under zero initial conditions. Since F is strictly discrete postive (real), it follows that H is strictly discrete bounded (real). Furthermore, since $4\Re\{p^*(m,n)q(m,n)\} = |u(m,n)|^2 - |y(m,n)|^2$ we have if $E_N \leq K_1N + K_2$ then

$$S_N = \sum_{T_N}\sum |u(m,n)|^2 - y|(m,n)|^2 \leq 4(K_1'N + K_2'). \tag{58}$$

It then follows from Theorem 4.1(i) that both $u|(m,n)|^2$ and $y|(m,n)|^2$ satisfy (13) for some constants $K_1', K_2', |\beta|$ independent of N such that $K_1' = |\beta|^2, K_2' = |\beta|^2K_2$:

$$\sum_{T_N}\sum |u(m,n)|^2 \leq K_1'N + K_2',$$

$$\sum_{T_N}\sum |y(m,n)|^2 \leq K_1'N + K_2'.$$

Thus, since $|p(m,n)|^2 + |q(m,n)|^2 = \frac{1}{2}(|u(m,n)|^2 = |y(m,n)|^2)$ it follows that

$$\sum_{T_N}\sum |p(m,n)|^2 \leq K_1'N + K_2',$$

$$\sum_{T_N}\sum |q(m,n)|^2 \le K_1'N + K_2'.$$

The rest of the arguement then follows as in the proof of Theorem 4.1.

5. A 2-D hyperstability based adaptive filtering scheme. In this section we demonstrate an application of 2-D hyperstability type results to problems of adaptive 2-D signal processing. The framework for this study is a direct 2-D generalization of that known as the HARF (hyperstable adaptive recursive filtering) algorithm in 1-D signal processing literature (see [28]–[32], [38]), which in turn is a modification of certain system identication techniques documented in [40].

The signal to tracked is assumed to be generated by a quarter plane causal linear shift invariant (LSI) ARMA process whose transfer function is assumed holomorphic in $\bar{D}$ and is thus stable in a strict sense [23]. Consequently,

$$y(m,n) = \sum_{\substack{k=0 \\ k+\ell\neq 0}}^{M_1}\sum_{\ell=0}^{N_1} a_{k\ell}\,y(m-k,n-\ell) + \sum_{k=0}^{M_2}\sum_{\ell=0}^{N_2} b_{k\ell}\,u(m-k,n-\ell) \tag{59}$$

The estimated signal $\hat{y}(m,n)$ is obtained as

$$\hat{y}(m,n) = \sum_{\substack{k=0 \\ k+\ell\neq 0}}^{M_1}\sum_{\ell=0}^{N_1} \hat{a}_{k\ell}(m,n)f(m-k,n-\ell) + \sum_{k=0}^{M_2}\sum_{\ell=0}^{N_2} \hat{b}_{k\ell}(m,n)u(m-k,n-\ell), \tag{60}$$

where $f(m-k,n-\ell)$, as described via (61), is the output from an auxiliary quarter plane shift varying recursive process driven by the same input $u(\cdot,\cdot)$ as in (59),(60):

$$\begin{aligned} f(m,n) &= \sum_{\substack{k=0 \\ k+\ell\neq 0}}^{M_1}\sum_{\ell=0}^{N_1} \hat{a}_{k\ell}(m+1,n+1)f(m-k,n-\ell) \\ &+ \sum_{k=0}^{M_2}\sum_{\ell=0}^{N_2} \hat{b}_{k\ell}(m+1,n+1)u(m-k,n-\ell) \end{aligned} \tag{61}$$

Finally, the coefficients $\hat{a}_{k\ell}(m,n)$ and $\hat{b}_{k\ell}(m,n)$ are adapted according to the following adaptation rules:

$$\hat{a}_{k\ell}(m+1,n+1) = \hat{a}_{k\ell}(m,n) + \mu_{k\ell}\hat{y}(m-k,n-\ell)\frac{\bar{v}(m,n)}{t(m,n)} \tag{62}$$

$$\hat{b}_{k\ell}(m+1,n+1) = \hat{b}_{k\ell}(m,n) + \rho_{k\ell}u(m-k,n-\ell)\frac{\bar{v}(m,n)}{t(m,n)}, \tag{63}$$

where

$$(64)\ \bar{v}(m,n) = y(m,n) - \hat{y}(m,n) + \sum_{\substack{k=0 \\ k+\ell\neq 0}}^{L_1} \sum_{\ell=0}^{L_2} c_{k\ell}[y(m-k,n-\ell) - f(m-k,n-\ell)]$$

and

$$(65)\quad t(m,n) = 1 + \sum_{\substack{k=0 \\ k+\ell\neq 0}}^{M_1} \sum_{\ell=0}^{N_1} \mu_{k\ell} f^2(m-k,n-\ell) + \sum_{k=0}^{M_2} \sum_{\ell=0}^{N_2} \rho_{k\ell} u^2(m-k,n-\ell),$$

where $\mu_{k\ell} \geq 0$, $\rho_{k\ell} \geq 0$, are some constants and with $c_{k\ell}$'s to be specified appropriately so that the algorithm converges, i.e., $y(m,n) \to \hat{y}(m,n)$ in some sense.

In fact, the adaptation scheme (60)–(65) can be more transparently viewed by considering the difference signal

$$(66)\qquad e(m,n) = y(m,n) - f(m,n),$$

which drives the moving average process with transfer function:

$$(67)\qquad \sum_{k=0}^{L_1} \sum_{\ell=0}^{L_2} c_{k\ell} z_1^k z_2^\ell;\ c_{00} = 1$$

to yield the averaged output $v(m,n)$. The estimation error signal $(y(m,n) - \hat{y}(m,n))$ is then added to it to yield $v(m,n) + (y(m,n) - \hat{y}(m,n))$, which in turn drives the adaptation algorithm. The scheme can be represented via signal flow diagram as in Fig. 3.

It is not necessary for the sizes of the masks $\{\hat{a}_{k\ell}(\cdot,\cdot)\}$ and $\{\hat{b}_{k\ell}(\cdot,\cdot)\}$ to be the same as those for $\{a_{k\ell}\}$ and $\{b_{k\ell}\}$. The only restriction is that the supports of the masks $\{\hat{a}_{k\ell}(\cdot,\cdot)\}$ and $\{\hat{b}_{k\ell}(\cdot,\cdot)\}$, respectively contain those of $\{a_{k\ell}\}$ and $\{b_{k\ell}\}$. However, to keep notational complications down we have assumed the masks to have same support.

We next demonstrate that (59)–(65) combined with adequate boundary conditions in fact determine a valid quarter plane causal recursive scheme. For this, we first assume that all of the boundary values needed for computation, namely $f(m,n)$, where m and/or n is negative and $\hat{a}_{k\ell}(m,0)$, $\hat{b}_{k\ell}(0,n)$ for $m \geq 0$, $n \geq 0$ are all zero. We further assume that $\hat{a}_{k\ell}(m,n)$, $\hat{b}_{k\ell}(m,n)$ are known for $(m,n) \in C_N$ and $(m,n) \in C_{N+1}$. Also, we assume that $f(m,n)$'s are known in $R_{P,N-1}$, where $P = 2\max(M_1,N_1)$ and the input $u(m,n)$'s are known in $R_{Q,N}$, where $Q = 2\max(M_2,N_2)$.

We then first use (60) to compute $\hat{y}(m,n)$ on C_N. Next we compute $\hat{a}_{k\ell}(m,n)$, $\hat{b}_{k\ell}(m,n)$ on C_{N+2} by using (62)–(65). Finally, $f(m,n)$ on C_N can be computed using (61). We then obviously have $f(m,n)$ in $R_{P,N}$, while $u(m,n)$ in $R_{Q,N+1}$ is obtained from the input data. The recursive scheme described above can then be repeated for larger values of N.

Our next task is to show the convergence of the estimated signal $\hat{y}(m,n)$ to the desired signal $y(m,n)$. For this, let us define

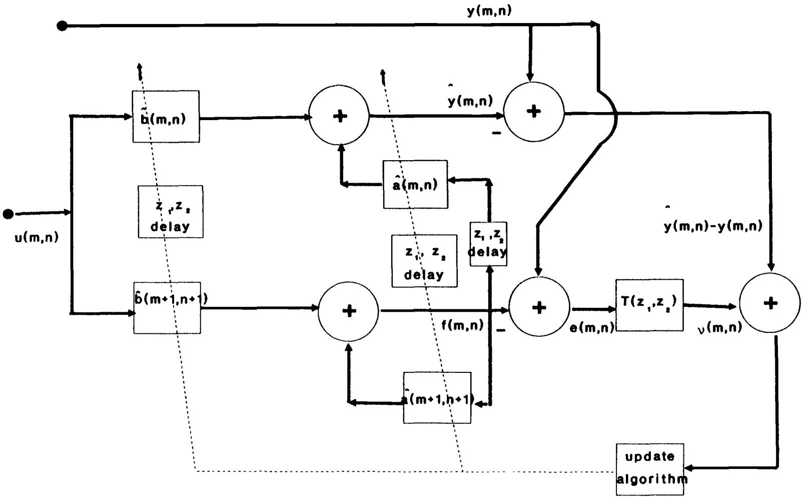


FIG. 3

$$
\begin{aligned}
&(68) \qquad w(m,n) = \sum_{\substack{k=0\\k+\ell\neq 0}}^{M_1}\sum_{\ell=0}^{N_1}[a_{k\ell} - \hat{a}_{k\ell}(m+1,n+1)]f(m-k,n-\ell)\\
&(69) \qquad\qquad\quad + \sum_{k=0}^{M_2}\sum_{\ell=0}^{N_2}[b_{k\ell} - \hat{b}_{k\ell}(m+1,n+1)]u(m-k,n-\ell).
\end{aligned}
$$

We obviously have (see Fig. 3) that

$$
v(m,n) = \sum_{k=0}^{L_1}\sum_{\ell=0}^{L_2} c_{k\ell}e(m-k,n-\ell). \tag{70}
$$

Furthermore, via some algebraic manipulations we can write

$$
e(m,n) = \sum_{\substack{k=0\\k+\ell\neq 0}}^{M_1}\sum_{\ell=0}^{N_1} a_{k\ell}e(m-k,n-\ell) + w(m,n). \tag{71}
$$

Consequently, the transfer function between $v(m,n)$ and $w(m,n)$ is given by

$$
T = \frac{\sum_{k=0}^{L_1}\sum_{\ell=0}^{L_2} c_{k\ell}z_1^k z_2^\ell}{1 - \sum_{k=0}^{M_1}\sum_{\ell=0}^{N_1} a_{k\ell}z_1^k z_2^\ell}. \tag{72}
$$

The entire feedback system of (62)–(65) can then be combined into the closed-loop configuration in Fig. 4., where the upper block is a linear shift invariant 2-D system with transfer function $T = T(z_1, z_2)$.

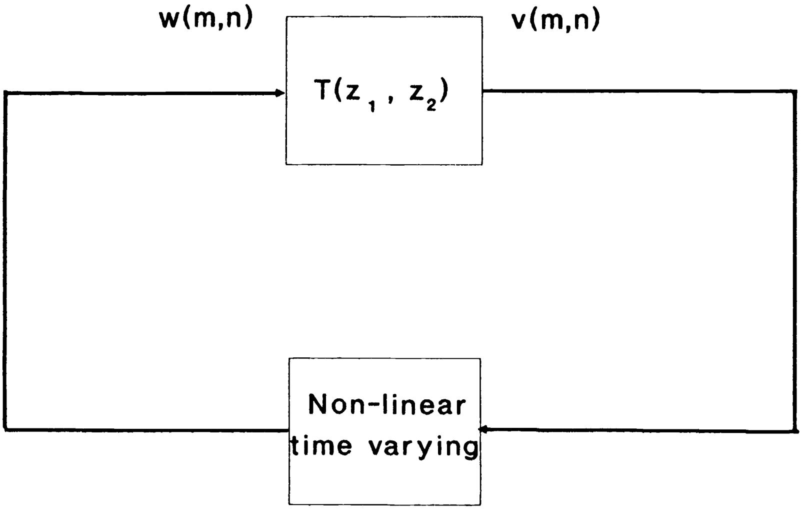


FIG. 4

Note that the denominator polynomial in T is the denominator of the transfer function of the unknown process given in (59), and is assumed to be devoid of zeros in $\bar{D}$, or in the terminology of [23] is a strictly Schur, thus, in particular, an immittance Schur polynomial. As shown in Theorem 27 of [23], given (real) $a_{k\ell}$'s it is then possible to find $c_{k\ell}$'s in (72) such that T is a discrete positive (real) function. It then follows that one can also find $c_{k\ell}$'s such that T is strictly discrete positive (real) (for example, if H is a (real) rational discrete positive, holomorphic in $\bar{D}$ then $1+H$ is (real) rational strictly discrete positive (real) and has the same denominator as in H).

Remark 5.1. We further wish to remark that the above considerations demand that the $a_{k\ell}$'s be known, which in reality are not. However, some knowledge of the values of the parameters $a_{k\ell}$'s, namely, a range of allowable values may be available. It has recently been shown [21] that via an extension of Kharitonov-theory for robustness of stability (see [43] for 1-D results) that in such a case it is possible to specify admissible intervals for the $c_{k\ell}$'s so that T in (72) remains strictly positive (real).

Before a complete proof of convergence of the adaptation algorithm can be given it would be necessary to focus on a 2-D state space realization of $T(z_1, z_2)$. The entire purpose of this discussion in the present context is to demonstrate that there is a realization of T for which it is reasonable to assume that the boundary conditions are zero.

For this, define the following variables for the state-variables of the Roesser's state space model:

$$(73) \qquad \bar{e}(m,n) \overset{\text{def}}{=} y(m,n) - f(m,n)$$

$$e_v(m,n) \overset{\text{def}}{=} \begin{bmatrix} \bar{e}(m,n-1) \\ \bar{e}(m,n-2) \\ \vdots \\ \bar{e}(m,n-N_1) \end{bmatrix} \tag{74}$$

$$e_h(m,n) \overset{\text{def}}{=} \begin{bmatrix} \bar{e}_{m-1,n} \\ \bar{e}_{m-2,n} \\ \vdots \\ \bar{e}_{m-1,N_1} \end{bmatrix}, \tag{75}$$

where

$$\bar{e}_{\mu,\nu} \overset{\text{def}}{=} \begin{bmatrix} \bar{e}(\mu,\nu) \\ \bar{e}(\mu,\nu-1) \\ \vdots \\ \bar{e}(\mu,\nu-N_1) \end{bmatrix}. \tag{76}$$

The feedback matrix $\mathcal{A}$ in the Roesser state space model is then defined as the matrix $\bar{A}$ from which the N_1-the row and the first column has been deleted and

$$\bar{A} \overset{\text{def}}{=} \begin{bmatrix} A_0 & A_1 & \cdots & \cdots & A_{M_1} \\ A_0 & A_1 & \cdots & \cdots & A_{M_1} \\ \bigcirc & I & \cdots & \cdots & \bigcirc \\ \vdots & \vdots & & & \vdots \\ \bigcirc & \bigcirc & \cdots & \cdots & I \end{bmatrix}, \tag{77}$$

where

$$A_i \overset{\text{def}}{=} \left.\begin{bmatrix} a_{i0} & a_{i1} & \cdots & a_{iN_1} \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}\right\} (n_2+1); \text{ for } i = 1 \text{ to } M_1 \tag{78}$$

and A_0 is the top companion matrix whose first row is $[\ a_{00} \quad a_{01} \quad \cdots \quad a_{1N_1}\ ]$.

The input matrix is $\mathcal{B} = [1\ 0 \cdots 0]^T$ and the output matrix $\mathcal{C}$ is the matrix $\bar{C}$ from which the first entry has been deleted and where

$$\bar{C}^T \overset{\text{def}}{=} [\ \bar{C}_0^T \quad \bar{C}_1^T \quad \cdots \quad \bar{C}_{M_1}^T\ ] \tag{79}$$

with

$$\bar{C}_i^T \overset{\text{def}}{=} [c_{i0}\ c_{i1} \cdots c_{iN_1}] \text{ for } i = 0 \text{ to } M_1. \tag{80}$$

We then have equations (81) and (82) for a 2-D Roesser state space model in standard form.

$$\begin{bmatrix} e_v(m,n+1) \\ e_h(m+1,n) \end{bmatrix} = \mathcal{A} \begin{bmatrix} e_v(m,n) \\ e_h(m,n) \end{bmatrix} + \mathcal{B}w(m,n) \tag{81}$$

$$v(m,n) = \mathcal{C} \begin{bmatrix} e_v(m,n) \\ e_h(m,n) \end{bmatrix} + w(m,n) \tag{82}$$

realizing the transfer function $T = T(z_1, z_2)$. The details of algebraic manipulations verifying this fact is routine and runs parallel to the 1-D case [29]. The essential point for us is that since it is reasonable to assume that $y(k,0) = y(0,\ell) = 0$ for $k < 0$, $\ell < 0$, by imposing

$$f(k,0) = f(0,\ell) = 0 \text{ for } k < 0, \ell < 0 \tag{83}$$

it can be seen in view of (73) to (76) that $e_v(0,n) = 0$, $e_h(m,0) = 0$ for all nonnegative values of m and n.

We are now ready to derive the promised convergence result.

THEOREM 5.1. *Assume the following regarding the above recursive scheme*:

1. *T in* (72) *is strictly discrete positive (real).*
2. *Boundary conditions in the auxiliary process f(m,n) are set to zero as in* (83).
3. *The input signal is bounded, i.e.,* $|u(m,n)| < K < \infty$ *for all* m, n; K *=constant.*
4. *Each of the estimated parameter values* $\hat{a}_{k\ell}(m,n)$, $\hat{b}_{k\ell}(m,n)$ *are zero along the boundary, i.e., if* $m = 0$ *and/or* $n = 0$.

Then the estimated signal $\hat{y}(m,n)$ *converges to the desired signal* $y(m,n)$ *in the sense that* $(\hat{y}(m,n) - y(m,n))$ *converges to zero in the almost everywhere sense.*

Proof. We first define the parameter estimation errors:

$$\bar{a}_{k\ell}(m,n) = a_{k\ell} - \hat{a}_{k\ell}(m,n) \tag{84}$$

$$\bar{b}_{k\ell}(m,n) = b_{k\ell} - \hat{b}_{k\ell}(m,n). \tag{85}$$

Also, routine algebraic manipulation with (62)–(66) and (70) yields

$$\bar{v}(m,n) = v(m,n)t(m,n). \tag{86}$$

Further manipulations as elaborated in [28] in the 1-D case can be performed with (86) to produce the following expression for $w(m,n)v(m,n)$:

$$\begin{aligned} 2w(m,n)v(m,n) &= s(m,n) - s(m+1,n+1) \\ &- \{\sum_{\substack{k=0 \\ k+\ell\neq 0}}^{M_1} \sum_{\ell=0}^{N_1} \mu_{k\ell} f^2(m-k,n-\ell)v^2(m,n) \\ &+ \sum_{k=0}^{M_2} \sum_{\ell=0}^{N_2} \rho_{k\ell} u^2(m-k,n-\ell)v^2(m,n)\}, \end{aligned} \tag{87}$$

where

$$s(m,n) = \sum_{\substack{k=0\\ k+\ell\neq 0}}^{M_1}\sum_{\ell=0}^{N_1} \bar{a}_{k\ell}^2(m,n)/\mu_{k\ell} + \sum_{k=0}^{M_2}\sum_{\ell=0}^{N_2} \bar{b}_{k\ell}^2(m,n)/\rho_{k\ell}. \tag{88}$$

Since $\mu_{k\ell} \geq 0$, $\rho_{k\ell} \geq 0$ by choice of these parameters, we have from (87) that for all (m,n):

$$2w(m,n)v(m,n) \leq s(m,n) - s(m+1,n+1). \tag{89}$$

Considering (89) for each $(m,n) \in T_N$, adding the resulting set of inequalities, and finally cancelling the common terms with opposite sign we then obtain (90)

$$2\sum\sum_{T_N} w(m,n)v(m,n) \leq \{s(0,0) + \sum_{m=1}^{N} s(m,0) + \sum_{n=1}^{N} s(0,n)\} - \sum_{D_N} s(m,n), \tag{90}$$

where $D_N = C'_{N+1} \cup C'_{N+2}$ and $C'_N = C_N \setminus \{(N,0)\} \cup \{(0,N)\}$.

Since as is obvious from (88) the last term in (90) can be dropped without affecting the inequality. Inequality (91) is then obtained by substituting for $s(m,n)$ from (88) in the resulting inequality:

$$\begin{aligned} &\sum\sum_{T_N} w(m,n)v(m,n) \\ &\leq \frac{1}{2}[\sum_{\substack{k=0\\ k+\ell\neq 0}}^{M_1}\sum_{\ell=0}^{N_1}\{\sum_{n=1}^{N} \bar{a}_{k\ell}^2(0,n) + \sum_{m=1}^{N} \bar{a}_{k\ell}^2(m,0) + \bar{a}_{k\ell}^2(0,0)\} \\ &\quad + \sum_{k=0}^{M_2}\sum_{\ell=0}^{N_2}\{\sum_{n=1}^{N} \bar{b}_{k\ell}^2(0,n) + \sum_{m=1}^{N} \bar{b}_{k\ell}^2(m,0) + \bar{b}_{k\ell}^2(0,0)\}]. \end{aligned} \tag{91}$$

Also, since the parameter estimates along the boundary $(m,0)$ and $(0,n)$ are set to zero, we have $\bar{a}_{k\ell}(0,n) = \bar{a}_{k\ell}(m,0) = a_{k\ell}$, $\bar{b}_{k\ell}(0,n) = \bar{b}_{k\ell}(m,0) = b_{k\ell}$. Thus, (91) above yields

$$\sum\sum_{T_N} w(m,n)v(m,n) \leq K_1 N + K_2, \tag{92}$$

where

$$K_1 = 2K_2 = \sum_{\substack{k=0\\ k+\ell\neq 0}}^{M_1}\sum_{\ell=0}^{N_1} a_{k\ell}^2 + \sum_{k=0}^{M_2}\sum_{\ell=0}^{N_2} b_{k\ell}^2. \tag{93}$$

Next, observe that $w(m,n)$ and $v(m,n)$ are, respectively, the input and output from the transfer function T, which is strictly discrete positive (real). Furthermore, since the system realizing T can be assumed to be operating under zero boundary conditions the total response due to the system (including the transient response) is equal to its zero state response and is solely determined by the transfer function T. Thus, the realization given by (81) and (82) can be replaced by a passive realization of T operating under zero boundary conditions without altering $w(m,n)$ and $v(m,n)$. Theorem 4.10 thus applies to the present situation, which in turn yields that

$$\sum_{T_N}\sum |v(m,n)|^2 \leq K_1'N + K_2' \tag{94}$$

for some constants K_1' and K_2'.

Next, observe that the moving average transfer function $\sum_{k=0}^{L_1}\sum_{\ell=0}^{L_2} c_{k\ell} z_1^k z_2^\ell$ is nonzero in $\bar{D}$, has $e(m,n)$ at its input and $v(m,n)$ satisfying (94) as its output when implemented with zero boundary conditions. Thus by invoking Lemma 4.9 it follows that

$$\sum_{T_N}\sum |e(m,n)|^2 \leq K_1''N + K_2'', \tag{95}$$

where K_1'' and K_2'' are some constants independent of N. We, therefore, have from Theorem 3.1 that $e(m,n) = y(m,n) - f(m,n)$ converges to zero almost everywhere. In particular, $e(m,n)$ is asymptotically bounded almost everywhere. But since $y(m,n)$ is the output from a quarter plane causal strictly discrete positive real, thus bounded input bounded output stable [15] transfer function with bounded input $|u(m,n)| < K$, $y(m,n)$ must be bounded for all $m, n \geq 0$. Consequently, it follows from Proposition 3.4(i) that $f(m,n) = y(m,n) - e(m,n)$ must be asymptotically bounded almost everywhere.

Next, recall that (86) can, by using (65), be explicitly written as

$$\begin{aligned}\bar{v}(m,n) &= v(m,n)\\ &\cdot\left\{1 + \sum_{\substack{k=0\\ k+\ell\neq 0}}^{M_1}\sum_{\ell=0}^{N_1}\mu_{k\ell} f^2(m-k,n-\ell) + \sum_{k=0}^{M_2}\sum_{\ell=0}^{N_2}\rho_{k\ell}u^2(m-k,n-\ell)\right\}.\end{aligned} \tag{96}$$

Invoking almost everywhere asymptotic boundedness property of $f(m,n)$, the boundedness $|u(m,n)| < K$, and Proposition 3.4 it follows that the term inside the brackets in (96) is asymptotically bounded almost everywhere. Since $v(m,n)$ converges to zero almost everywhere, by Proposition 3.3 so does $\bar{v}(m,n)$. Finally, by recalling from (64) that

$$\begin{aligned}&\hat{y}(m,n) - y(m,n)\\ &\quad = \bar{v}(m,n) - \sum_{\substack{k=0\\ k+\ell\neq 0}}^{L_1}\sum_{\ell=0}^{L_2} c_{k\ell}\{y(m-k,n-\ell) - f(m-k,n-\ell)\}\end{aligned} \tag{97}$$

and using the facts that both $\bar{v}(m,n)$ and $\{y(m,n)-f(m,n)\}$ converges to zero almost everywhere along with Proposition 3.2 we conclude that $\hat{y}(m,n)-y(m,n)$ converges to zero almost everywhere, thus completing the proof of Theorem 5.1.

We also have the following more restrictive result.

THEOREM 5.2. *Assuming* (1) *through* (3) *in Theorem* 5.1 *to hold and that the parameter estimates are such that for each k, ℓ*

$$\sum_{m=0}^{N} \bar{a}_{k\ell}^2(m,0) < \infty; \sum_{n=0}^{N} \bar{a}_{k\ell}^2(0,n) < \infty \tag{98}$$

and

$$\sum_{m=0}^{N} \bar{b}_{k\ell}^2(m,0) < \infty; \sum_{n=0}^{N} \bar{b}_{k\ell}^2(0,n) < \infty \tag{99}$$

hold true in the recursive scheme described via (59)–(65), *we have that $\hat{y}(m,n)$ converges to $y(m,n)$ in the usual sense as $(m+n) \to \infty$.*

Proof. It follows from (91),(98), and (99) that under the conditions stated

$$\sum\sum_{T_N} w(m,n)v(m,n) < \infty \tag{100}$$

for every N, i.e., (92) is satisfied with $K_1 = 0$ and some K_2, not necessarily given by (94). Following a similar chain of arguments we then conclude:

$$\sum\sum_{T_N} |e(m,n)|^2 < \infty$$

as a counterpart of (95). Thus, we have $e(m,n) \to 0$ as $(m+n) \to \infty$. The rest of the proof follows by mimicking arguments in the proof of Theorem 5.1, but now using convergence and boundedness in the usual sense.

Satisfaction of conditions (98) and (99), in view of (84) and (85), require that the chosen boundary values of the parameter estimates along the boundary $(m,0)$ and $(0,n)$ are asypmtotically (i.e., as $m \to \infty$, $n \to \infty$) the correct (unbiased) estimates. Although there is no way to guarantee this, some intelligent choice of these may, in practice, yield better results than the most conservative strategy of setting them to zero as in Theorem 5.1. A suggestion is to set $\hat{a}_{k\ell}(m,0) = \hat{a}_{k\ell}(m-1,1)$ and $\hat{a}_{k\ell}(0,n) = \hat{a}_{k\ell}(1,n-1)$, the latter quantities having been computed via the parameter update scheme (62) to (65). Similarly for the $\hat{b}_{k\ell}(m,n)$'s when m and/or n is zero. However, such a scheme, although reasonable from an intuitive viewpoint, defies our convergence analysis presented above.

6. Discussions and Conclusions. A continuous domain analogue of the 2-D hyperstability theory can be developed and the results applied to a continuous version of the adaptive filtering problem. Note that a continuous version of hyperstable adaptive identifier exists in 1-D literature [40].

A straightforward n-D ($n > 2$) extension of the theory developed here does not seem to be possible because of the following two reasons. First, a higher dimensional analogue of Landau's theorem does not exist, and second, perhaps more serious, is the infeasibility of synthesis of lossless n-D ($n > 2$) paraunitary matrices holomorphic in $\bar{D}$ [26]. Although a passive realizability theory based approach to higher dimensional problems is not possible due to the above reasons, since the essential contents of Theorems 4.1 and 4.10 do not state anything regarding the realization of $H(z_1, z_2)$ or $F(z_1, z_2)$, but only have to do with its input-output properties, namely, strict positivity or boundedness, an approach using exclusively transform domain methods may be fruitful.

A matricial generalization of the hyperstability Theorems 4.1 and 4.10 for multi-input–multi-output 2-D systems may be conjectured, but a proof would require a matrix version of Landau's theorem, or equivalently, an appropriate 2-D generalization of Youla's spectral factorability result [27] in our context.

The vital importance of the spectral factorability [27] result mentioned above in diverse areas of 1-D system theory is well known. Since Landau's theorem can, in some sense, be viewed as a 2-D scalar analogue of this result, we believe that many other ramifications of it would be possible in the context of 2-D systems. In the present paper we have demonstrated its use in 2-D hyperstability theory and in 2-D adaptive filtering only.

In all of the 2-D recursions considered in the present paper, the set of points C_N of the 2-D lattice space for a fixed N, has been viewed as a "computational wavefront," i.e., all points on C_N can be simultaneously computed from C_{N-1}, C_{N-2}, $\cdots$, etc. While this is attractive and may have some advantages in parallel implementation, the role of C_N as a computational wavefront can be replaced by L-shaped sets: $L_N = \{(m, N); m = 0 \text{ to } N\} \cup \{(N, n); n = 0 \text{ to } N\}$ in the 2-D lattice space. In fact, all results of §§3 and 4 can be modified accordingly. It should then be possible to derive an appropriate version of the HARF algorithm for this recursive scheme. Apart from other possibilities of recursion, the 2-D HARF can be simplified as in 1-D [30], [31] and it is likely that its properties can be studied via 2-D Lyapunov theory [20].

A. Some results on 2-D passive synthesis. THEOREM A.1 (E. LANDAU [7]). *Let $f(x_1, x_2)$ be a polynomial in x_1, x_2 with real coefficients, which is such that $f(x_1, x_2) \geq 0$ for all real x_1, x_2. Then $f(x_1, x_2)$ can be written as*

$$f(x_1, x_2) = \frac{1}{d(x_1)} \sum_{i=1}^{\nu} N_i^2(x_1, x_2) \tag{101}$$

where $d(x_1)$, $N_i(x_1, x_2)$; $i = 1$ to ν are polynomials with real coefficients and the integer ν is at most equal to 4.

LEMMA A.2. *Let $S = N/g$ be a bounded (real) matrix of size $(1 \times n)$ in two variables p_1, p_2 i.e., $SS_* \leq 1$ in* Re $p_1 > 0$, Re $p_2 > 0$, *where N is a polynomial row vector and g is the least common denominator of S. Then $(gg_* - NN_*)$ can be decomposed as:*

$$gg_* - NN_* = \frac{1}{dd_*} \sum_{i=1}^{m} P_{i*} P_i \tag{102}$$

where each P_i is a polynomial in two variables p_1 and p_2, and d is a one variable polynomial in p_1 (or p_2) devoid of zeros in the open right half plane and $m \leq 2$.

Proof. Since S is a bounded (real) function we have that $(gg_* - NN_*) \geq 0$ for all $p_1 = j\omega_1$ and $p_2 = j\omega_2$. Furthermore, for $p_1 = j\omega_1$ and $p_2 = j\omega_2$ $(gg_* - NN_*)$ can be viewed as a polynomial in ω_1 and ω_2, with real coefficients. Thus, by Landau's theorem we can write for $p_1 = j\omega_1$ and $p_2 = j\omega_2$ that

$$gg_* - NN_* = \frac{1}{D(\omega_1)} \sum_{i=1}^{\nu} P_i^2(\omega_1, \omega_2) \tag{103}$$

where D and each P_i are polynomials with real coefficients and $\nu \leq 4$.

Assume without loss of generality that ν =even, and let $m = \nu/2$. By using the identity $P^2 + Q^2 = (P + jQ)(P - jQ)$ in (103) we can then write

$$\sum_{i=1}^{\nu} P_i^2(\omega_1, \omega_2) = \sum_{i=1}^{m} M_i(\omega_1, \omega_2) M_i^*(\omega_1, \omega_2) \tag{104}$$

where M_i's in (104) are polynomials in ω_1, ω_2 with coefficients complex conjugates of those in M_i^*'s.

Note further that since $(gg_* - NN_*) \geq 0$ for all $p_1 = j\omega_1$, $p_2 = j\omega_2$ and real ω_1, ω_2 it follows from (103) that $D(\omega_1) \geq 0$ for all ω_1. Thus, by a standard result, $D(\omega_1)$ can be written as $D(\omega_1) = M_D(\omega_1) M_D^*(\omega_1)$, where $M_D(\omega_1)$ is a polynomial with coefficients complex conjugates of those in $M_D^*(\omega_1)$. To summarize, the polynomial identity

$$M_D(\omega_1) M_D^*(\omega_1)(gg_* - NN_*) = \sum_{i=1}^{m} M_i(\omega_1, \omega_2) M_i^*(\omega_1, \omega_2) \tag{105}$$

holds for all real valued ω_1, ω_2.

Now consider the polynomials $P_i = P_i(p_1, p_2)$, of which the coefficients are such that $P_i(j\omega_1, j\omega_2) = M_i(\omega_1, \omega_2)$. Then $P_{i*}(j\omega_1, j\omega_2) = M_i^*(\omega_1, \omega_2)$. Thus, we have $P_i(p_1, p_2) P_{i*}(p_1, p_2) = M_i(\omega_1, \omega_2) M_i^*(\omega_1, \omega_2)$ for $p_1 = j\omega_1$, $p_2 = j\omega_2$, where ω_1, ω_2 are arbitrary real. Via the same argument we can then write: $d_1(p_1) d_{1*}(p_1) = M_D(\omega_1) M_D^*(\omega_1)$, where $d_1(p_1)$ is a real polynomial in p_1 and $d_{1*}(p_1)$ is its paraconjugate. Thus, from (105) we can write that

$$d_1 d_{1*}(gg_* - NN_*) = \sum_{i=1}^{m} P_i P_{i*} \tag{106}$$

holds for $p_1 = j\omega_1$ and $p_2 = j\omega_2$, with ω_1, ω_2 arbitrary real, where the criptic notation $d_1 = d_1(p_1)$, $P_i = P_i(p_1, p_2)$ have been used. Since (106) is a polynomial equation we can conclude from an analytic continuation argument that (106), in fact, holds for all p_1 and p_2.

Finally, we can write $d_1 d_{1*} = dd_*$ via a regrouping of the factors of $d_1 d_{1*}$ such that the polynomial $d = d(p_1)$ does not have any zero in open right half plane. The validity of (102) is thus established from (106). The form (102) with $d = d(p_2)$ is obtained in exactly similar manner by choosing D to be a polynomial in ω_2 in the Landau decomposition.

THEOREM A.3 (Continuous embedding). [2] *Let N/g be a 2-D (real) rational matrix of size $(1 \times n)$, where $N = N(p_1, p_2)$ is a polynomial row vector of size $(1 \times n)$ and $g = g(p_1, p_2)$ the least common denominator polynomial of N/g. If N/g is, in addition, a bounded (real) matrix then there exists a 2-D lossless bounded (real) rational $(m \times m)$ matrix $(m > n)$ (i.e., H is holomorphic in* Re $p_1 > 0$, Re$p_2 > 0$ *with $HH_* = I_m$) such that H can be partitioned as*

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{21} & H_{22} \end{bmatrix}; \text{ where } H_{11} = \frac{N}{g}. \tag{107}$$

Remark A.1. An analogous result, if N/g is an $(n \times 1)$ vector, can be proven in a similar manner, but we do not require this in the context of this paper. However, if N/g is an $(m \times n)$ matrix $(m, n > 1)$ then a similar result can be conjectured, but is presently not available.

Proof of Theorem A.3. Define an $(n \times n)$ polynomial matrix S_{11} such that the first row of S_{11} is N/g and the rest of the elements are zero. Also, since N/g is discrete bounded (real) it follows from Lemma A.2 that there exists a polynomial row vector $V = V(p_1, p_2)$, say, of size $(1 \times m)$ such that

$$1 - \frac{NN_*}{gg_*} = \frac{VV_*}{(gd)(gd)_*}, \tag{108}$$

where d is a polynomial in p_1 only, devoid of zeros in the open right half plane. Next, define a rational matrix S_{12} of size $n \times (m + n - 1)$ as

$$S_{12} = \begin{bmatrix} \frac{V}{gd} & 0 \\ 0 & I_{n-1} \end{bmatrix} \tag{109}$$

and two further matrices S_{21} and S_{22} as

$$S_{21} = S'_{21}\frac{\alpha}{\alpha_*}; \; S'_{21} = -S_{12*}(I + S_{11*})^{-1}(I + S_{11}) \tag{110}$$

$$S_{22} = S'_{22}\frac{\alpha}{\alpha_*}; \; S'_{22} = I - S_{12*}(I + S_{11*})^{-1}S_{12}, \tag{111}$$

where α is a polynomial to be specified later in course of the proof, and I's are identity matrices of appropriate sizes. It can then be routinely verified that the square matrix S of size $(m + 2n - 1)$ as in

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix} \tag{112}$$

satisfies $SS_* = I$.

[2] Subsequent to the preparation of the paper, this result has been derived from Landau's theorem for matrix (not necessarily vector) valued N/g in [17] and its consequences in passive synthesis is discussed in [18].

We next claim that by proper choice of α it is possible to make S holomorphic in Re $p_1 > 0$, Re $p_2 > 0$. For this, first note that S_{11} and S_{12} are clearly holomorphic in Re$p_1 > 0$, Re $p_2 > 0$, and choose α to be product of those irreducible factors in the least common denominator of $S_{12*}(I + S_{11*})^{-1}$ having a zero in Re $p_1 > 0$, Re $p_2 > 0$. Since S_{11} is a discrete bounded matrix, $(I - S_{11})(I + S_{11})^{-1}$ is a positive matrix. Thus, it follows from the identity

$$(I + S_{11})^{-1} = \frac{1}{2}\{I + (I - S_{11})(I + S_{11})^{-1}\} \tag{113}$$

that $(I + S_{11})^{-1}$ is a positive matrix. Consequently [22], $(I + S_{11})^{-1}$ and thus, $(I + S_{11})^{-1}S_{12}$ is holomorphic in Re $p_1 > 0$, Re$p_2 > 0$, which in turn imply that the least common denominator of $S_{12*}(I + S_{11*})^{-1}$ does not have a zero in Re$p_1 < 0$, Re$p_2 < 0$. Since the chosen α is a factor of this last mentioned denominator, α does not have any zero in Re$p_1 < 0$, Re$p_2 < 0$ either. Thus, the factor α/α_* is holomorphic in Re$p_1 > 0$, Re$p_2 > 0$, which additionally cancels all factors having zeros in Re$p_1 > 0$, Re$p_2 > 0$ in the denominators of S'_{21}, S'_{22}. We have thus constructed a lossless bounded (real) S of which S_{11} and, thus, $H_{11} = N/g$ is a submatrix at its top left corner. Finally, H as in (107) is obtained by appropriately repartitioning S.

Remark A.2. Note that the above result does not imply that gH = polynomial matrix, nor does it imply that the H obtained via the procedure outlined above has the smallest possible size for a given N/g. Thus, if the discrete lossless bounded H is synthesized as in [24,25,26] to yield a synthesis of N/g, the minimality of neither the dynamic elements, i.e., the number of p_1 and p_2 type elements (in fact, this number is at most $\deg_1(dg)+\deg_2(dg)$ in the present case) nor the number of "fully absorbing" ports (in continous domain these correspond to resistors) created in the synthesis procedure is ensured.

THEOREM A.4 (Discrete embedding). *Let B/a be a discrete bounded (real) rational matrix function in two variable z_1, z_2, where B is a polynomial row vector of size $(1 \times n)$ and a is the least common denominator polynomial of the entries of B/a. Then there exists a discrete lossless bounded (real) rational matrix G (i.e., G is holomorphic in $\bar{D}$ and $\tilde{G}G = I$) such that G can be partitioned as*

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}; \textit{ where } G_{11} = \frac{B}{a}. \tag{114}$$

Proof. Consider the action of the double bilinear transformation

$$z_i = \frac{1 - p_i}{1 + p_i}; \ i = 1, 2 \tag{115}$$

on $G_{11} = B/a$, which produces the bounded (real) rational matrix N/g of the variables p_1, p_2, in which N is a polynomial vector of size $(1 \times n)$ and g is the least common denominator polynomial of the entries of N/g, i.e., we have

$$\frac{N}{g} = \left[\frac{B}{a}\right]_{z_i = \frac{1-p_i}{1+p_i}}; \ \frac{B}{a} = \left[\frac{N}{g}\right]_{p_i = \frac{1-z_i}{1+z_i}}; \ i = 1, 2. \tag{116}$$

Invoking Theorem A.3 we obtain the lossless bounded (real) rational matrix H as in (107) with $H_{11} = N/g$. Now consider the action of the inverse transform to yield

$$G = [H]_{p_i = \frac{1-z_i}{1+z_i}}\ ;\ i = 1, 2. \tag{117}$$

We then correspondingly also have

$$\frac{B}{a} = G_{11} = [H_{11}]_{p_i = \frac{1-z_i}{1+z_i}}\ ;\ i = 1, 2$$

Since it can be trivially shown from the corresponding property of H that G is a discrete lossless bounded (real) rational matrix, the proof of the present theorem is complete.

Acknowledgments. The author expresses deepest gratitude to Professor A. Fettweis for continued discussions on passive system theory over the last several years, to Dr. A. Kummert of his group for pointing out reference [7] and its potential relevance in scalar continuous embedding in the summer of 1988, and to Dr. A. Tan (a former graduate student) for checking many of the calculations reported in this paper. He also thanks Professor Christian Berg for pointing out [9], and an anonymous reviewer for constructive suggestions.

REFERENCES

[1] V. M. Popov, *Hyperstability of Control Systems*, Springer-Verlag, New York, 1973.

[2] P. Caines, *Linear Stochastic Systems*, John Wiley, New York, 1989.

[3] C. Desoer, *On the relation between pseudopassivity and hyperstability*, IEEE Trans. Circuits and Systems, CAS-22(1975), pp. 897–898.

[4] B. D. O. Anderson, *A simplified viewpoint of hyperstability*, IEEE Trans. Automat. Control, 13(1968), pp. 292–294.

[5] B. D. O. Anderson and S. Vongpanitlerd, *Network Analysis and Synthesis*, Prentice–Hall, Englewood Cliffs, NJ, 1973.

[6] J. C. Willems, *Least squares stationary optimal control and the algebraic Riccati equation*, IEEE Trans. Automatic Control, 39(1987), pp.427–429.

[7] E. Landau, *Über die Darstellung definiter Funktionen durch Quadrate*, Math. Ann., 62(1906), p. 272.

[8] D. Hilbert, *Über Ternäre Definite Formen*, Acta Math., 32(1893), pp. 169–197.

[9] T. Y. Lam, *The algebraic theory of quadratic forms*, Benjamin, Reading, MA, 1973.

[10] B. L. van der Waerden, *Modern Algebra*, Vol. 2, Third edition, Frederick Unger Publishing Co., 1950.

[11] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer-Verlag, New York, Berlin, 1971.

[12] V. A. Yakubovich, *A frequency theorem for the case in which the state and control spaces are Hilbert spaces, with an application to some problems in the synthesis of optimal controls, Part* I, Siberian Math. J., 15(1974), pp. 639–668; *Part* II, Siberian Math J., 16(1975), pp. 1081–1102.

[13] D. Z. Arov, *Passive Linear stationary dynamic systems*, Siberian Math. J., 20(1979), pp. 211–228. 211-228,

[14] B. Sz-Nagy and C. Foias, *Analysis of Operators on Hilbert space*, North–Holland, Amsterdam, 1970.

[15] N. K. Bose, *Applied Multidimensional Systems Theory*, Van Nostrand Reinhold, New York, 1982.

[16] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, University of California Press, 1958.

[17] A. Kummert, *Spectral factorization of two-variable parahermitian polynomial matrices*, preprint.

[18] ———, *The synthesis of two-dimensional passive n-ports containing lumped elements*, preprint.

[19] S. Y. Kung, B. C. Levy, M. Morf, and T. Kailath, *New results in* 2-D *systems theory, part* II: 2-D *state-space models, realization and the notions of controllability, observability, and minimality*, Proc. IEEE, 65(1977), pp. 945–961.

[20] B. D. O. Anderson, P. Agathoklis, E. I. Jury, and M. Mansour, *Stability and the matrix Lyapunov equation for discrete* 2*-dimensional systems*, IEEE Trans Circuits and Systems, CAS-33 (1986), pp. 261–267.

[21] S. Basu, *On boundary implications of stability and positivity properties of multidimensional systems*, Proc. IEEE, Special Issue on Multidimensional Signal Processing, 76 (1990), pp. 614–626.

[22] A. Fettweis and S. Basu, *New results on stable multidimensional polynomials–part* I: *Continuous case*, IEEE Trans. Circuits and Systems, 34(1987), pp. 1221–1232.

[23] S. Basu and A. Fettweis, *New results on stable multidimensional polynomials–part* II: *Discerete case*, IEEE Trans. Circuits and Systems, 34(1987), pp. 1264–1274.

[24] D. C. Youla, *The synthesis of networks containing lumped and distributed elements*, Proc. Symp. Generalized Networks, Polytechnich Inst. Brooklyn Press, New York, 1966.

[25] T. Koga, *Synthesis of finite passive n-ports with prescribed positive real matrices of several variables*, IEEE Trans. Circuit Theory, 15(1968), pp. 2–23.

[26] A. Kummert, *Beiträge zur Synthese Mehrdimensionaler Reactanzmehrtore*, Ph.D. thesis, Lehrstuht f ür Nachrichtentechnik, Ruhr Universit ät, Bochum, West Germany, 1988.

[27] sc D. C. Youla, *On the factorization of rational matrices*, IEEE Trans. Inform. Theory, 18(1961), pp. 172–189.

[28] C. R. Johnson, Jr., *A convergence proof for a hyperstable adaptive recursive filter*, IEEE Trans. Inform. Theory, IT-25(1979), pp. 745–749.

[29] ———, *Another use of the Lin–Narendra error model*: HARF, IEEE Trans. Automat. Control, AC-25(1980), pp. 985–988.

[30] M. G. Larimore, J. R. Treichler, and C. R. Johnson, Jr., SHARF: *An algorithm for adapting IIR digital filters*, IEEE Trans. Acoust. Speech Signal Process., ASSP-28(1980), pp. 480–440.

[31] C. R. Johnson, Jr., M. G. Larimore, J. R. Treichler, and B. D. O. Anderson, SHARF *convergence properties* IEEE Trans. Acoust. Speech Signal Process., ASSP-29(1981), pp. 428–440, June 1981.

[32] C. R. Johnson, Jr., *Adaptive IIR filtering*: *current results and open issues*, IEEE Trans. Inform. Theory, IT-30(1984), pp. 237–250.

[33] J. J. Shynk, *Adaptive IIR filtering*, IEEE ASSP Magazine, April 1989, pp. 4–21.

[34] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice–Hall, Englewood Cliffs, NJ, 1985.

[35] S. Haykin, *Adaptive Filter Theory*, Prentice–Hall, Englewood Cliffs, NJ, 1986.

[36] J. R. Treichler, C. R. Johnson, Jr., and M. G. Larimore, *Theory and Design of Adaptive Filters*, John Wiley, New York, 1987.

[37] M. L. Honig and D. G. Messerschmitt, *Adaptive Filter*, *Structures*, *Algorithms*, *and Applications*, Kluwer Academic Pubs., Norwell, MA, 1984.

[38] C. F. N. Cowan, and P.M. Grant, ed., *Adaptive Filters*, Prentice–Hall, Englewood Cliffs, NJ, 1985.

[39] B. D. O. Anderson et. al, *Stability of Adaptive Systems–Passivity and Averaging Analysis*, MIT Press, Cambridge, MA, 1986.

[40] I. D. Landau, *Adaptive Control*: *The Model Reference Approach*, Marcel Dekker, New York, 1979.

[41] G. Goodwin and K. S. Sin, *Adaptive Filtering, Prediction and Control*, Prentice–Hall, Englewood Cliffs, NJ, 1984.

[42] S. Sastry and M. Bodson, *Adaptive Control*, *Stabililty*, *Convergence*, *Robustness*, Prentice–Hall, Englewood Cliffs, NJ, 1989.

[43] S. Dasgupta, *Kharitonov like theorem for systems under passive nonlinear feedback*, Proc. of 26th conf. on Decision and Control, Los Angeles, CA, December 1987, pp. 2062-2063.

[44] M. M. Hadhoud and D. W. Thomas, *The two-dimensional adaptive* LMS (TDLMS) *algorithm*, IEEE Trans. Circuits and Systems, Vol. 35(1988), pp. 485–494.

[45] P. Chan and J. S. Lim, *One-Dimensional processing for adaptive image processing*, IEEE Trans. Acoust. Speech Signal Process., ASSP-33(1985), pp. 117–125.

[46] S. T. Alexander and S. A. Rajala, *Image compression results using the* LMS *adaptive algorithm*, IEEE Trans. Acoust. Speech Signal Process., Vol. ASSP-33(1985), pp. 712–714.

[47] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, National Bureau of Standards, Gaithersburg, MD, 1964.